Agentschap voor Innovatie door Wetenschap en Technologie
IWT
SBO Security and Privacy for Online Social Networks

Document type	Report
Title	D5.2- Report on Research Activities
Work Package	WP5
Deliverable Number	D5.2
Editor(s)	Bettina Berendt, KU Leuven
Dissemination level	Public
Preparation date	27 November 2014
Version	1.0

The IWT SBO SPION Project

No.	Participant name	Country	Department	Participant role
1	KU Leuven	BE	COSIC/ESAT	Coordinator
2	KU Leuven	BE	DISTRINET	Partner
3	KU Leuven	BE	DTAI	Partner
4	KU Leuven	BE	ICRI	Partner
5	Vrije Universiteit Brussel	BE	SMIT	Partner
6	Universiteit Gent	BE	OWK	Partner
7	Carnegie Mellon University	USA	Heinz	Partner

Contributors

	Name	**Organisation**
1	Bettina Berendt	KU Leuven
2	Bo Gao	KU Leuven
3	Seda Gürses	KU Leuven
4	Thomas Peetz	KU Leuven
5	Jo Pierson	Vrije Universiteit Brussel

External Contributors

	Name	**Organisation**
1	Ciham Demir	Helmut-Schmidt Gymnasium Hamburg
2	Gebhard Dettmar	Helmut-Schmidt Gymnasium Hamburg
3	Silvia Gabrielli	CREATE-NET Italy
4	Anthony Jameson	German Research Center for Artificial Intelligence / DFKI Saarbrücken
5	Allison Littlejohn	Glasgow Caledonian University
6	Anoush Margaryan	Glasgow Caledonian University
7	Anthony Morton	University College London
8	Sören Preibusch	Microsoft Cambridge
9	Riina Vuorikari	EUN Schoolnet

Table of Contents

Keyword List

Privacy-preserving data mining methods in social networks and other environments; Privacy awareness methods in social networks and other environments; Privacy awareness and data mining in social networks; Discrimination-aware data mining methods; Privacy decision making

1 Executive Summary

According to the Description of Work, the objective of Deliverable 5.2 is to report on DTAI's research activities in months M25–M46. Specifically, the Deliverable consists of seven published papers that investigate goals and decision making based on the feedback and awareness created by F&A tools (see Deliverable D5.1),

1. in situations in which the "correct" decision is comparatively clear (or given),

– and investigating the less clear-cut situations:

2. when there is no clearly right ("good") or wrong decision
3. when the measure of what is "good" is also the performance target
4. when different stakeholders have different notions of "good"

To study goals and decision making, we performed conceptual analyses, developed and built software tools, and carried out user studies. An introduction summarises and contextualises these seven papers.

This Deliverable was planned, in the Project Proposal, as consisting of five conference-level papers. The actual version consists of one conference paper, three standard journal papers, a journal paper of book length, and two book chapters. One of the journal papers is a collaboration of researchers from three SPION partners: DTAI, ESAT and SMIT. One of the journal papers and one of the book chapters are a collaboration of researchers from the SPION partner DTAI and a SPION user partner.

2 Introduction: Privacy, feedback, awareness, goals and decision making

In our previous summary-of-work Deliverable (2), we introduced the notion of privacy feedback and awareness (PFA) tools. We defined these by contrasting them with two established classes of privacy-enhancing technology: cryptographic tools and access control. Briefly put, PFA tools aim to raise awareness and maybe also change behaviour by giving feedback on past and possible current/future actions of the user and other stakeholders as well as on the effects of these actions on privacy-related outcomes. To the extent that users are in control of information flows involving their personal data, the aim of PFA tools is to help them do this in a "good" way. Privacy is regarded as involving withdrawal/concealing as well as connecting/disclosing information.

In years 1 and 2 of SPION, we concentrated on the notion of awareness (and on what feedback to give to raise it) itself. In years 3 and 4, we focussed on what awareness is supposed to support: better decision making relative to relevant goals.[1] We investigated the notion of "*better* decision making" itself and in relation to how FA tools can support it,

1. in situations in which the "correct" decision is comparatively clear (or given),

– and investigating the less clear-cut situations:

2. when there is no clearly right ("good") or wrong decision
3. when the surrogate measure of what is "good" is also the performance target
4. when different stakeholders have different notions of "good"

2.1 Decision making when the "good" decision is assumed to be known or given

2.1.1 FreeBu and the task of audience management

First, we have consolidated the work on the FreeBu tool for friend grouping and contextualisation, started in (2, Section 2.3.1).

The second version of FreeBu (PAPER1; PAPER2) was designed based on the results of the first round of evaluation (joint work with SMIT, currently under revision). It has also been enriched by now providing for a multi-perspective view on one's Facebook friends corresponding to multiple strategies of grouping them. The important role of multiple perspectives for learning (via, it may be argued, awareness) was pointed out in our earlier research in SPION, see (2, PAPER 2).

The choice of grouping method/algorithms was based on an algorithmic evaluation and a user study. The algorithmic evaluation compared the algorithmically generated groups to a reference dataset of "ground truth groups" created by users in response to the request to sort their online contacts into groups. In the user study, participants were asked to think of three posts that they would not like all of their friends to see, and then the performance of FreeBu in supporting them in configuring access control for these posts was measured, and compared to Facebook's Smart Lists. Thus, this setting reflects the motivation for FA tools explained in the introduction of (2): to help users (a) reflect on the circumstances in which visibility of social-media content becomes an

[1] We assume this instrumental role for awareness for the areas of privacy and fairness/non-discrimination that we have dealt with in this project. It is also implicit in the discussion, in the wider HCI awareness literature, of how to measure awareness when this cannot be done directly: "changes in performance on tasks for which good situational awareness is essential", see (2, p. 15). Of course, there may be other areas in which awareness is an end in itself.

issue and (b) consider the consequences of using access-control mechanisms ("privacy settings"). For enhancing ecological validity, study participants were asked to think of posts individually; for experimental control, their task (or: a "good" privacy-related choice) was given to them: to restrict visibility by using the FreeBu tool. In further work (see Section 2.4), we have argued and shown that this is a rather restricting and unrealistic assumption, and advocated a more user-driven view of "tasks" (and therefore, a fortiori, of what a "good decision" is).

2.1.2 Feedback and awareness: white box vs. black box in decision support

We have also extended our work on exploratory discrimination-aware data mining (DADM). Here too, it appears to be relatively straightforward what a "good" decision is: one that does not discriminate, at least in the sense of avoiding unlawful discrimination (for example, in a banking or insurance context).

Based on our earlier work on DCUBE-GUI (2, Section 2.3.2), we have provided a conceptual analysis of exploratory DADM, grounded DADM in and contrasted it with legal and sociological notions of (anti-)discrimination, and carried out a large-scale experimental user study of the relative merits of constraint-oriented and exploratory DADM (cDADM resp. eDADM) for different use cases. This is described in (PAPER3).

Results showed that the discrimination-aware tool support in the eDADM and cDADM treatments led to significantly higher proportions of correct decisions, which were also motivated more accurately. There is significant evidence that the relative advantage of discrimination-aware techniques depends on their intended usage. For users focussed on making and motivating their decisions in non-discriminatory ways, cDADM resulted in more accurate and less discriminatory results than eDADM. For users focussed on monitoring for preventing discriminatory decisions and motivating these conclusions, eDADM yielded more accurate results than cDADM.

These results indicate that the white-box approach of exploratory DADM that also characterises feedback and awareness tools, can lead to more transparent decision making than the black-box approach of constraint-oriented DADM, whose algorithmic "sanitization" of undesired data-mining results may lead to an illusory sense of safety in decision making.

The results however also show that it is less straightforward to say – or in any case, to capture informatically – what a "good", non-discriminating, socially fair decision is. This problem was suspended for the study described in the following section, but taken up again in the publications of Sections 2.3 and in particular 2.4.

2.2 Decision making when there is no clearly right ("good") or wrong decision

Privacy-related decisions are instances of a larger class of decisions: those in which there is no clear-cut "right" or "wrong" choice. Interestingly, while the field of HCI has produced substantial research and results about how to design user interfaces that support choices where there is a clear "right" and "wrong" (e.g., a button that does implement a certain functionality vs. many other widgets that do not), *preferential choices* (i.e. those where this is not clear-cut) have been much less studied.

(PAPER4) attempts to close this gap. It presents a comprehensive framework for in-depth analyses of preferential choice and decision making in HCI: how people make such choices at all, and how interface design can support (or impede) them in such choice strategies. In addition to many illustrative examples from a wide range of areas, we chose privacy-related decisions and online communities as two subfields that we analysed in detail. One of the interfaces/systems analysed with respect to how it supports privacy-related decisions is FreeBu (see Section 2.1.1 above) – but of course, the wider view of this general publication also shows how many more open issues there are in supporting users in privacy-related decisions, on online social networks and elsewhere.

2.3 Decision making when the measure of what is "good" is also the performance target

Some of the PFA tools that we analysed in the SPION State-of-the-Art Deliverable (1) are "analytics" tools: They show one or more measures of online behaviour to the user, under the assumption that these numbers are indicators of some relevant latent variable, and that certain values represent "good" privacy behaviour and outcomes. Straightforward examples are tools with a traffic-light metaphor: The assumed goal is "good privacy", measured by how much of one's profile and interaction information is visible to others (the less, the better); this is fed back to the user through a numerical score and/or colour (green = good, red = bad). The assumption is that to minimize deviation from the goal, users will change their privacy settings towards less visibility.

This is based on a straightforward cybernetic metaphor of systems with a feedback loop, such as a thermostat that has a goal (desired temperature), a measured state (actual temperature), and that effects cooling or heating in order to minimize the discrepancy (1). It is also based on the application of this cybernetic idea in "key performance indicators" systems that have been used to support business management for a long time.

"Analytics", "indicators", or "dashboard" tools and interfaces may be argued to be the dominant form in which data analyses are today presented to non-expert users in all kinds of areas under the assumption that this raises transparency, creates incentives, improves behaviour and decisions, etc. Examples include individual health, sports, and other activities' measures ("quantified self") and the organisation and management of cities ("city dashboards"). The question is whether this model of decision making works.

There are still too few such systems and evaluations in the area of privacy (see also (1; 2)). We therefore analysed an application area in which "analytics" tools, including those that can be regarded as FA tools, are more common: learning analytics. In (PAPER5), we study learning analytics with respect to continuing professional education. We present a case study of learning analytics for our user partner EUN Schoolnet's eTwinning (the European social networking platform for teachers[2]) and analyse potentials and limitations of this decision architecture.

Three of our conclusions are of particular relevance for privacy awareness: First, analytics tools are always also surveillance tools, which themselves may cause privacy violations and/or hamper the free expression of users (here: learners) and thereby have negative effects on outcomes. Second, people are no thermostats: they respond to measures they see portrayed as specifying a target value, and they change their behaviour (e.g. by gaming the system). This is widely known in finance, banking, and the social sciences in general, summarised succinctly as "When a measure becomes a target, it ceases to be a good measure."[3] A strict orientation towards one goal may therefore be both ineffective and too constraining - which is related to our findings on exploratory vs. constraint-oriented discrimination-aware data mining (Section 2.1.2). Third, analytics tools that involve the learners neither in design nor as users may actual disempower them, and we recommend a stronger integration of users in these processes. We have built especially on the last two observations in the work reported next.

There are analytics tools that are at the same time FA tools (where the data subject herself sees the measures) and others that are not (where somebody else than the data subject sees the measures). Similarly, there are FA tools that are at the same time analytics tools (where the tool shows some measure to be monitored and controlled in the expectation of it leading to better privacy or other outcomes) and others that are not (where the tool invites exploration and reflection in a less

[2] www.etwinning.net

[3] This is known as, among other names, "Campbell's Law", explained by the author in full as "The more any quantitative social indicator (or even some qualitative indicator) is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." (Campbell, D. T. (1976). *Assessing the Impact of Planned Social Change.* The Public Affairs Center, Dartmouth College, Hanover New Hampshire, USA.)

directed way). Both FreeBu and our proposals for exploratory discrimination-aware data mining focus on exploration and in that sense belong to the latter class; although DCUBE-GUI's scores and the red/green flagging in the user study of eDADM do suggest clear measures of success. In the following section, we describe our first steps into building on the lessons learned from studying the problems of analytics tools; continuing these efforts is one of the areas of our future work beyond this project.

2.4 Decision making when different stakeholders have different notions of "good"

In the previous sections, we have elaborated on the problems that may arise when simplistic notions of "good" decisions and outcomes are used. In the final parts of this Deliverable, we describe current work on dealing with these challenges.

2.4.1 FreeBu and its user-perceived affordances

The first step in considering different stakeholders is to become aware that the designer of a tool or user study is also a stakeholder, no more but also no less. Thus, the (privacy) goals given in a study (e.g. "choose the people who should see a certain post with the help of this tool") and/or suggested by a tool (e.g. "make contents visible to as few people as possible") may be just one stakeholder's goals. FreeBu was expressly designed to support users' exploration; the question is what affordances they perceive in this tool.[4] We explicitly frame it in this way because we cannot claim to have access to users' "real, innate" privacy goals – too many factors, including the tool itself and the uses it suggests, will also influence what users *say* they want from a tool or what they *do* when interacting with a tool. We concluded that the best we can achieve is to gain complementary perspectives on their goals with a mixed-methods design.

With respect to the FreeBu and its support for friend grouping and audience management (Section 2.1.1), we have carried out two mixed-methods user studies for eliciting users' own perceptions of the tool's value and affordances. The results show that grouping (and thus audience management), but also reflection on who one's friends actually are, and to a lesser extent "de-friending" (as one of the more radical measures of audience management) are the major perceptions of what FreeBu is "good" for. We also obtained results on user preferences for visualization types and on the influence of visual variables on tool use. This study was joint work with SMIT and is currently under review.

2.4.2 "Tool clinics" and teaching about privacy

Going beyond specific tools, in (PAPER6) we have proposed "tool clinics" as a method for eliciting not only users', but various stakeholders' notions of what a privacy tool or research artefact should do, i.e. what would be "good" decisions and outcomes of design. Our plan is to test the method sketched in that paper in teaching settings or at conferences.

Finally, (PAPER7) marks one of the foci of research and practical work with which we finish the SPION project: the embedding of tools and other informatics-centric and short-term "solutions" into more long-term and thereby hopefully ultimately more viable strategies. Specifically, (PAPER7) describes a result of an ongoing user-partner collaboration: a lesson series on privacy for high schools. The lesson series is highly interdisciplinary, combining informatics, economics and politics/law. At two key junctures in the series, pupils are asked to perform a role play. The goal is to recognise

[4] By focusing on user perceptions, we use Norman's rather than Gibson's notion of affordances, see Gibson, J.J. (1977). The Theory of Affordances. In Robert Shaw & John Bransford (Eds.), *Perceiving, Acting, and Knowing*. Lawrence Erlbaum Associates; Norman, D. (1988). *The Psychology (*later: *Design of Everyday Things*. New York: Basic Books.

conflicts of interest between the users and providers of online social networks, between citizens, companies, and society, and between different fundamental rights. In this way, we aim to help pupils understand the multifaceted nature of privacy, its complex role in and for democracy, and recognize their own options.

The paper also describes the experiences of the first complete run[5] of the series from the perspectives of the teacher and one of the pupils. We explicitly chose this predominantly qualitative approach as an exploratory, formative evaluation. The outcome is thus a carefully optimistic conclusion: the series is generally viewed as relevant and interesting, such teaching can increase knowledge and create indignation about certain data-processing activities that may lead to privacy violations and discrimination. We regard this indignation as a first step towards more consistent changes in attitudes. However, it remains difficult to change behaviour. (This mirrors some of the findings of (4) that summarises a large part of the work done by SPION's OWK partner. It should be noted that, in contrast to (4), our work in (PAPER7) focuses on the practice of teaching rather than on educational science.) We need more runs of such lesson series, and given that even "grown-ups" and whole societies struggle with what their privacy-related attitudes and behaviours should be, why should we expect it to be easy to teach it?

2.5 (Some) conclusions and future work

We wish to thank our colleagues for the highly inspiring collaboration over these four years, and IWT for the funding, flexibility and freedom that made this work possible.

As any good research effort, SPION has not only allowed us to answer some questions – it has led the way to many new ones. In particular, it has provided us with many insights into the necessity to think of feedback and awareness at a systemic and long-term level extending beyond a software-centric approach. SPION has paved the way for ongoing and projected work that aims at developing broader systemic methods by aiming both at a closer integration of informatics, software development, professional ethics, and teaching strategies, and a continuing critical examination of the potentials and limits of this approach and how it can be "good" for privacy.

2.6 A note on authorship: Collaborations in and through SPION

The papers collected in this Deliverable reflect the highly collaborative nature of the SPION project and at the same time integrated external partners from the areas of privacy decision making:

- Publications (PAPER1; PAPER2) are SPION DTAI work. The second version of FreeBu described there was developed in close collaboration with SPION colleagues Ralf De Wolf and Jo Pierson (VUB-SMIT), in particular through the user studies described in Section 2.4.1.
- Publication (PAPER3) continues the external collaboration with Sören Preibusch that was reported in (2). We are very grateful to SPION colleague Brendan Van Alsenoy (KUL-ICRI) for many fruitful discussions and valuable feedback.
- Publication (PAPER6) is joint work with SPION colleagues Seda Gürses (KUL-ESAT) and Jo Pierson (VUB-SMIT), as well as external contributor Anthony Morton. The paper is the result of a working group at the Dagstuhl seminar *'My Life, Shared' - Trust and Privacy in the Age of Ubiquitous Experience Sharing* (July 2013), which was co-organised by SPION colleague Alessandro Acquisti (CMU).
- Publication (PAPER5) is joint work with SPION user partner Riina Vuorikari (EUN Schoolnet).

[5] In addition, an abridged version was run in a different class at the same school

- Publication (PAPER7) is joint work with SPION user partners Gebhard Dettmar and Cihan Demir (Helmut-Schmidt-Gymnasium Hamburg). This lesson series was inspired by text leading towards DTAI's contribution to the SPION Privacy Manual (3), it was worked out and deployed at a school in Hamburg twice, it was the subject of presentations and teacher trainings that the two first authors of (PAPER7) have done, and it has been the basis for follow-up work outside the context of SPION.
- The idea for the collaboration towards publication (PAPER4) arose after the keynote given by Anthony Jameson at the SPION-DTAI Workshop "Privacy feedback and awareness – the what, the how and the who".[6]

2.7 A note on the complete respectively partial inclusion of publications into this Deliverable

(PAPER4) is a complete book, and the text contribution of its SPION co-author Bettina Berendt concentrated on the section on privacy-related decision making. We have therefore included only this section, prepended by the introduction that is needed to put this section into the context of the framework. For the same reason, we have included only Anthony Jameson (the first author of the whole book) and Silvia Gabrielli (with whom we collaborated on the privacy section) as external contributors to this Deliverable. The full paper is hyperlinked.

(PAPER7) is in German. For the purposes of this Flemish/English project Deliverable, we have therefore included only the title page (as an illustration), the current state of the English-language online summary of our collaboration, an English-language overview of the lesson series, and a Dutch-language presentation of the lesson series (presented at the SPION Educational Workshop[7]). The full paper is hyperlinked.

[6] http://people.cs.kuleuven.be/~bettina.berendt/SPION/workshop_index.html, 16 April 2013

[7] 13 February 2014

Bibliography

[A. Published papers that are part of the present Deliverable]

[PAPER1] Bo Gao and Bettina Berendt. Circles, posts and privacy in egocentric social networks: An exploratory visualization approach. In *ASONAM, Niagara Falls, Canada, 25-28 August 2013, 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 792–796. IEEE Computer Science Press, 2013.

[PAPER2] Bo Gao and Bettina Berendt. Friends and circles – a design study for contact management in egocentric online social networks. In Jalal Kawash, editor, *Online Social Media Analysis and Visualization*, Lecture Notes in Social Networks. Springer-Verlag, 2014.

[PAPER3] Bettina Berendt and Sören Preibusch. Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. *Artificial Intelligence and Law*, 22(2):175–209, 2014.

[PAPER4] A. Jameson, B. Berendt, S. Gabrielli, C. Gena, F. Cena, F. Vernero, and K. Reinecke. Choice architecture for human-computer interaction. *Foundations and Trends in Human-Computer Interaction*, 7(1-2):1–235, 2013.

[PAPER5] Bettina Berendt, Riina Vuorikari, Allison Littlejohn, and Anoush Margaryan. Learning analytics and their application in technology-enhanced professional learning. In Allison Littlejohn and Anoush Margaryan, editors, *Technology-enhanced Professional Learning. Processes, Practices and Tools*, pages 147–157. Routledge Taylor & Francis Group, 2014.

[PAPER6] Anthony Morton, Bettina Berendt, Seda Gürses, and Jo Pierson. “tool clinics" - embracing multiple perspectives in privacy research and privacy-sensitive design. *Dagstuhl Reports*, 3(7):96–104, 2013.

[PAPER7] B. Berendt, G. Dettmar, C. Demir, and T. Peetz. Kostenlos ist nicht kostenfrei. oder: If you’re not paying for it, you are the product. *LOG IN*, 178/179:41–56, 2014.

[B. Further SPION references]

[1] A. Acquisti, E. Balsa, B. Berendt, D. Clarke, R. De Wolf, C. Diaz, B. Gao, S.F. Gürses, A. Kuczerawy, J. Pierson, F. Piessens, R. Sayaf, T. Schellens, F. Stutzman, B. Van Alsenoy, and E. Vanderhoven. SPION Deliverable 2.1 State of the Art, 2011. COSIC Internal Technical Report, K.U. Leuven, Belgium. http://www.cosic.esat.kuleuven.be/publications/article-2077.pdf

[2] B. Berendt, D. Clarke, R. De Wolf, B. Gao, J. Pierson, S. Preibusch, and R. Sayaf. SPION Deliverable 5.1 Report on Research Activities (Identity Management), 2012. COSIC Internal Technical Report, K.U. Leuven, Belgium. http://www.cosic.esat.kuleuven.be/publications/article-2302.pdf.

[3] Brendan van Alsenoy, Frank Piessens, Ero Balsa, Seda Gürses, Bettina Berendt, Bo Gao, Willem De Groef, Martin Valcke, Bram Lievens, Ralf De Wolf, Fred Stutzman, Eva Lievens, Dave Clarke, Rula Sayaf, Claudia Diaz, Bart Preneel, Tammy Schellens, Jef Ausloos, Thomas Peetz, Ellen Vanderhoven, Jo Pierson, Alessandro Acquisti, and Jos Dumortier. SPION Deliverable 9.2.2 Privacy Manual, 2012. COSIC Internal Technical Report, K.U. Leuven, Belgium. http://www.cosic.esat.kuleuven.be/publications/article-2298.pdf.

[4] Ellen Vanderhoven. *Raising risk awareness and changing unsafe behaviour on social network sites: A design-based research in secondary education*. PhD thesis, University of Ghent, Belgium, 2014. http://users.ugent.be/~mvalcke/CV/Doctoraat_EllenVanderhoven.pdf.

3

PAPER 1:

Circles, posts and privacy in egocentric social networks: An exploratory visualization approac

Circles, Posts and Privacy in Egocentric Social Networks: An Exploratory Visualization Approach

Bo Gao
Department of Computer Science
KULeuven
Telephone: +32 (0)16/32.77.00
Email: firstname.lastname@cs.kuleuven.be

Bettina Berendt
Department of Computer Science
KULeuven
Telephone: +32 (0)16/32.77.00
Email: firstname.lastname@cs.kuleuven.be

***Abstract*—The users in Online Social Networks (OSN) may share private information with wrong friends. One approach to tackle this issue is by applying community discovery methods in egocentric networks to automatically generate friend circles for the user. There is however a discrepancy between the predicted circles and the circles that the user has in mind. A deep rooted reason is that it only makes sense when the circles are considered under certain usage. We designed and implemented an exploratory visualization tool that can help users determine the visibilities of their online posts. More specifically, we first examined the state-of-the-art community discovery methods for egocentric networks, then proposed a new visualization design with fine-grained control for the user to interact with the circles and make visibility decisions. Finally, we conducted an experimental user study evaluating the usefulness of this design.**

***Keywords*—Online Social Networks; Visualization; Circles; Design; Privacy**

I. Introduction

An Online Social Network (OSN) today can hold hundreds of millions of users[1], such as Facebook. Large amount of online personal information is exchanged daily. This phenomenon has raised privacy concerns. Two types of such concerns can be distinguished: *social* and *instrumental* [1]. *Social privacy* concerns how and when personal information is shared with others within an OSN (e.g. [2], [3], [4]), whereas *instrumental privacy* concerns the personal data access by service providers, governments or other corporations (e.g. [5], [6]). In this paper, we focus on *social privacy*. More specifically, we are interested in the tools that help users control the flow of personal information shared with friends in Egocentric OSN (EOSN). An EOSN is a network with the vertices representing people and the edges representing certain relationships among them. It is centered on one person whom we call the ego. The friends of the ego, whom we call the alters, must be directly linked to the ego. An alter can also connect to other alters.

As previous studies have suggested [7], [8], [2], [9], in order to manage the personal information flow, it is important for the user to categorize the friends into circles, lists or communities[2]. The community discovery algorithms may help users in this regard. However, as elaborated in Section II, there is a discrepancy between the predicted circles and the circles that the user has in mind. This calls for a type of application that can help its users effectively utilize the output of a community discovery algorithm. In this paper, we introduce one such application.

The contributions of this paper are: First, an exploratory tool is described. The tool is to help its users categorize friends more effectively in EOSN. Second, we describe an experimental user study to evaluate the effectiveness of the circles when a user makes visibility decisions about posts. Third, a new kind of interactive visualization was designed to assist in fine-grained exploration of hierarchical circles.

The structure of this paper is as follows: In Section II, we motivate our design choices by reviewing related works. Section III describes the design of the tool. Section IV gives an account of our user study for evaluating the tool. In Section V, we conclude by a discussion of future work and a summary of the paper.

II. Related Works and Design Choices

A. Notation

We denote an EOSN as a graph $G = (V, E, F)$, in which V is a set of vertices, with each vertex v an alter, usually labeled with a name. E is a set of edges with each edge $e = (u, v)$, with $u, v \in V$ representing a relation between u and v. For example, a relation can be formed if u and v are mutual friends in G or u follows v. F is a set of features describing V. A typical feature can be v's profile information, such as "gender is female". There exists a function assigning features to vertices, $\phi : V \times F \rightarrow \{f, v\}$. We denote an algorithm-predicted circle as c and a manual circle created by a user as $\tilde{c}$, with $c \subseteq V$, $\tilde{c} \subseteq V$. Correspondingly, the set of generated circles is denoted as C and user-created circles as $\tilde{C}$. We use p to denote a post. A post may include updating status, changing profile information, uploading/sharing photos/videos, tagging

[1] http://en.wikipedia.org/wiki/List_of_social_networking_websites

[2] We use "circle", "list" and "community" interchangeably in this paper. These words all refer to a collection of alters in an EOSN, usually with common characteristics. However, "community" is often used as a more general term in the field of community discovery algorithms, while "circle" and "list" are mentioned more in the EOSN context.

names in photos, liking, commenting, etc. A *Visibility Decision* refers to an ego's decision on the visibility of his post to each alter.

B. Community Discovery Algorithms

Community discovery in networks is a general problem and many algorithms exist [10], [11]. There are three categories of community discovery algorithms based on the types of input data — *Category 1* takes only the network E into account. The relationship of mutual friends or follower-followee forms an edge. In general, this category produces circles composed of densely connected alters. *Category 2* only considers the features F. This category produces circles composed of alters sharing common feature(s). *Category 3* makes use of both E and F. We are interested in the algorithms that may predict similar circles as the ones a user would manually create.

McAuley and Leskovec [12] examined eight community discovery algorithms from the above three categories and proposed a new model that outperforms the others. $Accuracy(c, \tilde{c})$ (Equation 1) is used to determine how well a set of predicted circles matches its manual counterpart. BER is short for Balanced Error Rate. The linear assignment between $c \in C$ and $\tilde{c} \in \tilde{C}$ is determined by the Munkres algorithm [13]. Let H be the set of pairs of circles that are matched. The average accuracy $Accuracy(C, \tilde{C})$ between the predicted circles C and the manual circles $\tilde{C}$ is shown in Equation 2.

$$\begin{aligned} Accuracy(c, \tilde{c}) &= 1 - BER(c, \tilde{c}) \\ with\ BER(c, \tilde{c}) &= \frac{1}{2}\left(\frac{|c \backslash \tilde{c}|}{|c|} + \frac{|\tilde{c} \backslash c|}{|\tilde{c}|}\right) \end{aligned} \tag{1}$$

$$Accuracy(C, \tilde{C}) = \frac{\sum_{(c,\tilde{c}) \in H} Accuracy(c, \tilde{c})}{min(|C|, |\tilde{C}|)} \tag{2}$$

For convenience, we name the model and the corresponding algorithm in [12] as GMF, short for "Generative Model for Friendships". GMF takes profile information to construct edge probabilities based on the EOSN network. The circles are then found by maximizing the overall probability. The number of circles needs to be pre-determined. GMF can also be computationally expensive — we ran it on the ten Facebook users' data provided in [12], it took more than an hour on average to generate circles for each user[3]. These limitations make us consider alternative algorithms for our tool. Another community discovery algorithm developed by Newman [14] is not among the eight baselines in [12]. It takes only the network data as input. The circles are found by maximizing the modularity of the network[4], and the number of circles is automatically determined. With the "Jmod" implementation of Newman's algorithm [15], the average computation time for each of the ten Facebook users is less than eight seconds. For simplicity, we refer to this algorithm as MOD. Figure 1 summarizes the performances in accuracy of the two algorithms running on the ten Facebook users' data. We see that MOD outperforms GMF with respect to the three K values. This suggests that modularity-based circles can be a good choice to be integrated in the tool design.

[3]The algorithm is run on a computer with i7-2600 (3.4GHz, 8MB cache) CPU and 16GB memory. The source code and the datasets can be found at the author's website: http://i.stanford.edu/~julian/.

[4]The edge density in a circle should be larger than that on average in the whole graph.

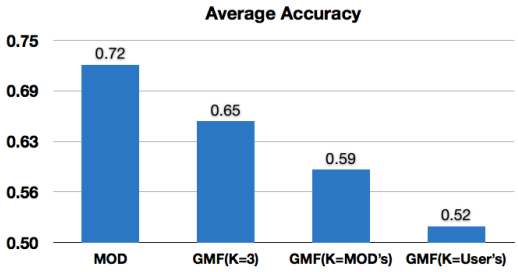


Fig. 1. Average accuracy scores of MOD and GMF on the ten Facebook users' data. The number of circles K for GMF is set to different values. They are K = 3, K is equal to that of MOD for each user and K is equal to that of each user's manual circles.

C. Discrepancy between Predicted and Manual Circles

Though a community discovery algorithm can predict reasonably good circles, it is unlikely that it can make a perfect prediction. This attributes to the fact that circle-creation is inherently subjective. In a labeling exercise [12], the manual circles were obtained by letting the users assign label(s) to describe their friends. The friends with the same label(s) are considered to be in the same circle(s). This encourages overlapping circles because users tend to assign multiple labels to a friend. In a card-sorting exercise [16], each friend's name is written on a card. Several cards were pre-selected and spread on a table. A participant is then asked to assign the rest of the cards to the pre-selected ones to form groups. In principle, the same friend can be assigned to different groups, but since people tend to assign a friend just once, overlaps are rare. We see that people create circles differently under different circumstances. Therefore, it is critical to enable the user to explore his friends based on an initial set of predicted circles that is "good enough", and adapt the circles for certain purpose. We may then evaluate the usefulness or effectiveness of these circles according to how well they have fulfilled that purpose. Exploration and adaptation of the circles by the user require an exploratory visualization approach.

D. Presentation and User Interaction of Circles

Major OSN sites such as Facebook, Google+ provide users with grouping functions. But except for Facebook smart lists, manual grouping has remained as the only way to organize and manage friends. It has also been quantitatively demonstrated that users' perceptions of their audience size do not match reality, since not enough feedback is provided for the users to be aware of the audience composition [17]. A visualization for EOSN circles was proposed [9]. It presents the composition of friends by labeling and resizing the circles. Our visualization addresses three improvements: 1) specifying the exact positioning of the circles to avoid overlapping layout; 2) specifying a way of browsing all the alters (members) in a circle. 3) enabling granular exploration of the circles. These points are particularly necessary given that the number of friends one might have in OSN is increasing, while empirical observations discourage displaying more than nine or ten items to be judged by a user [18], [19]. Moreover, current OSN lack the tools to let users manage the granular boundaries between multiple social groups as effectively as in their quotidian lives [20]. It thus becomes critical to provide users with a tool that enables granular exploration. This visualization design is detailed in

the next section.

III. The Tool Design

A. Modularity-based Community Discovery with Granularity

The original MOD algorithm (Section II) is non-hierarchical. The communities are discovered when further division does not lead to an increase of the modularity. For each derived community, we obtain a subgraph. The same algorithm may then be applied to each subgraph, deriving sub-communities. As such, we adapt the original algorithm into a hierarchical one. We refer to this modified algorithm as H-MOD. In the next subsection, we show how circles or sub-circles are divided with user-interactions. When we adopt the community discovery algorithm hierarchically, we make the visualization more fine-grained.

B. Exploratory Visualization of Circles

In this subsection, we introduce a new form of visualization. The circles are aligned and manipulatable by zooming and dragging. Their sizes are scaled to provide a visual order. To encourage exploration, we only provide the user with "zooming/panning" and automatic division to explore the circles [21], [22]. We use desaturated, sometimes adjacent colors to make the visualization more aesthetic [22]. This also promotes the tool's usability [23]. The details of the design are as follows:

The ego's circles are presented as in Figure 2 (left). We call the area where the circles are drawn the canvas. The grey dot in the center of the canvas represents the ego. The radii and positions of the circles are determined according to the number of people in each circle[5]. The circularly aligned grey dots represent the members of that circle. The lighter grey dot in the center, which we call the handle of that circle, represents the circle as a whole, labeled with a name. With the handle, the user can move the whole circle around and address all the members to make visibility decisions (Section IV). The curves linking the members and the handle provides the user visual cues of the belongingness of the members. The members within the circles with small radii are hidden from sight in order to display a clean, non-overlapping overview. Hidden members and their names can be brought to display with zooming. When the user zooms into one circle, newly generated sub-circles are presented if the subgraph corresponding to the circle is divisible. This is depicted in Figure 3. The user can also align all the names in a (sub)circle in a grid on the canvas (Figure 4).

IV. Experimental User Study

In this section, we describe the experimental user study that evaluates the effectiveness of the exploratory visualization tool for users' *Visibility Decisions*, with Facebook smart lists as our baseline[6]. Facebook smart lists detect communities based on the information about the user's education, work and current city. For example, the friends who went to the same school as the user are put into the same list.

There were 16 participants, 25-45 years old, from eight countries. Among them are Ph.D researchers, company employees and Master students. We divided the the participants

[5]For the detailed circle-positioning algorithm, see http://people.cs.kuleuven.be/~bo.gao/papers/ASONAM2013/GranularCircles_position_algo.pdf.

[6]https://www.facebook.com/help/204604196335128/

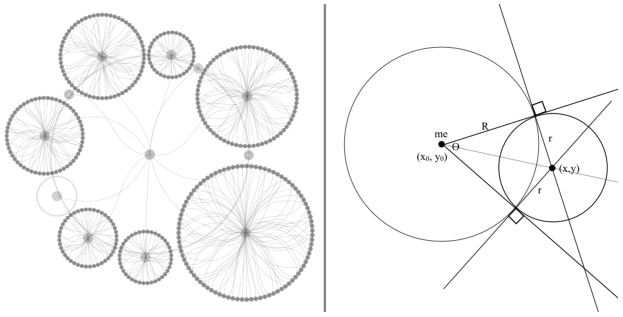


Fig. 2. Left: an overview of the circles' layout. Right: an illustration of drawing a circle around the ego.

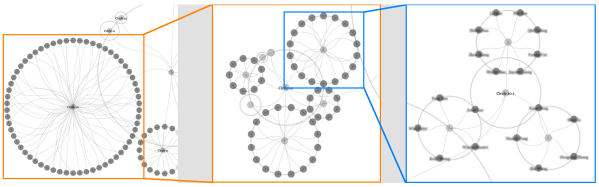


Fig. 3. An illustration of the hierarchical circles driven by a user's zooming. The names of the alters are blurred in this example to protect the user's privacy.

equally into two groups A and B. Group A used the tool we described in Section III. We took Group B as our baseline group. This group used the same visualization interface (Subsection III-B), but the underlying predicted circles were based on Facebook smart lists. The alters that were not in any smart list were put together into an extra circle. In this way, we removed the potential interference from using different interfaces. Our hypothesis is that users can make visibility decisions more effectively with the proposed exploratory visualization approach than the baseline approach. Each participant in both groups performed the following task comprised of two parts: elicitation of regrets in posts and visibility decision making.

Task: For the first part of the task, each participant was asked to identify his regretted posts. Though recent studies have investigated regrets in OSN from different aspects [24], [25], we chose to let our participants explicate their own regrets, because it is easier for a person to relate to his personal experience. A distinction was made between complete and partial regrets. A complete regret meant that the post was supposed to be seen by no one. A partial regret was where the participant did not mind his post being seen or intends his posts to be seen by some of the friends, but he failed to block the undesired friends. Since a complete regret entailed concealing the corresponding post completely, which would render a visibility decision trivial, we guided the participants to only think of partial regrets. Each participant was encouraged to think of three posts. A post needs to be specific enough to let the participant define its visibility to each friend. In total, 48 posts were collected. The types and the frequencies of the regretted posts are summarized in Table I. Note that some posts are of multiple types.

For the second part of the task, the participants were divided equally into two groups A and B. Each group has 8 participants and 24 posts. As shown in Figure 4, when a participant thinks an alter can see the post, he clicks on the dot, whose color turns from grey to blue. Clicking on the handle

TABLE I. PARTICIPANTS' REGRETTED POSTS

	Categories of Regretted Posts	Frequencies
a	I shouldn't express my bad mood or negative opinion.	6
b	I shouldn't ask for that advice or help.	1
c	There are uploaded photos depicting me in a way that I do not want to show to everyone.	15
d	language-specific posts	2
e	religious or political posts	5
f	I would have wanted to not show the post to that group of people for a particular reason.	6
g	I would have wanted to show the post only to that group of people for a particular reason.	6
h	inappropriate jokes	9

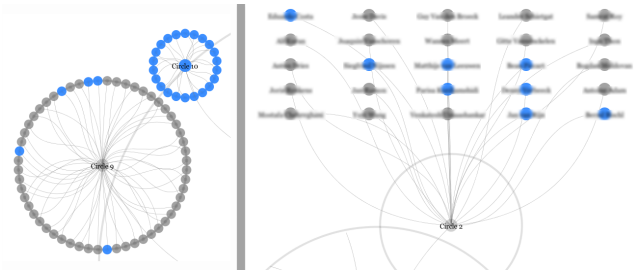


Fig. 4. By mouse-click, the participant can toggle individual members or a whole circle to indicate whether a post can be visible to them. An alter turns blue if the post is visible to him. Clicking the "handle" in the center of a circle toggles the whole circle (and its decedent circles if its hierarchical). On the right, we see that the members (labeled with their names) in a circle are aligned in a grid layout.

(the centered dot) of a circle makes the post visible to every member in that circle. The participants were allowed to work at their own pace until they are satisfied with their decisions.

Result: We use two measures to evaluate how effective the two approaches are for making visibility decisions: Accordance $Accordance(p) \in [0,1]$ (Equation 3) that calculates the average percentage of the members in a circle who can/cannot see the post p, and Entropy $Entropy(p) \in [0,1]$ (Equation 4) that calculates the overall information (in bits) needed to determine the whether a member in a circle can see a post. $Accordance(p) = 0$ or $Entropy(p) = 1$ means that on average, half a circle can see the post while the other half cannot, which means the set of circles is unhelpful. $Accordance(p) = 1$ or $Entropy(p) = 0$ means that on average, the members in the same circle have the same visibility status. That is, every circle, as a whole, can or cannot see the post, which is the case where the circles are fully utilized to make visibility decisions. The difference between the two measures is that the circles are treated equally in Accordance, while in Entropy, each circle is weighted according to the number of people in it, so that the visibility percentages in larger circles contribute more to the result.

The exploratory visualization interface is analogous to a binary-classification tree that tries to help the user utilize "pure" (sub)circles in terms of visibility decisions. The circles were firstly divided until they are indivisible according to the graph modularity or they are pure. We then used the leaves of the tree as the final set of circles [7]. Moreover, when there was only a small number of alters who could(not) see a post (e.g. less than five), Group A and B performed similarly well.

[7]The detailed division algorithm and an example can be found at http://people.cs.kuleuven.be/~bo.gao/papers/ASONAM2013/GranularCircles_division_algo.pdf

This is because a participant can simply handpick the people that he wants to target, any grouping solution becomes trivial. Let us denote the number of alters to whom a post is or is not visible, whichever smaller, as α. We call a visibility decision with $\alpha \geq \alpha_{th}$ an α_{th}-*Visibility Decision*. Thereby, grouping tools can be of more service to a user when α_{th} is larger. When we raise α_{th} to five, 38 posts out of 48 remain in the two groups, with 19 posts for each group. Figure 5 shows the Accordance and Entropy scores on average for Group A and B with $\alpha_{th} = 1$ and $\alpha_{th} = 5$ respectively.

$$
\begin{aligned}
&Accordance(p) = 2 \cdot (A_{show}(p) + A_{hide}(p)) - 1 \text{ with} \\
&A_{show}(p) = \left(\frac{\sum_{c \in C_v} N_{c,p}}{N}\right) \frac{\sum_{c \in C_v} \frac{N_{c,p}}{|c|}}{|C_v|} \text{ and} \\
&A_{hide}(p) = \left(\frac{\sum_{c \in C_{nv}} (|c| - N_{c,p})}{N}\right) \frac{\sum_{c \in C_{nv}} \frac{|c| - N_{c,p}}{|c|}}{|C_{nv}|}
\end{aligned} \tag{3}
$$

$$
\begin{aligned}
&Entropy(p) = \sum_{c \in C} \frac{|c|}{N} Entropy(c,p) \text{ with} \\
&Entropy(c,p) = -\frac{N_{c,p}}{|c|} \cdot log_2 \frac{N_{c,p}}{|c|} \\
&- \frac{|c| - N_{c,p}}{|c|} \cdot log_2 \frac{|c| - N_{c,p}}{|c|}
\end{aligned} \tag{4}
$$

C_v is the set of the circles containing members to whom p is visible. C_{nv} is the set of the circles containing members to whom p is not visible. Note that C_v and C_{nv} may overlap. $N_{c,p}$ is the number of the alters to whom p is visible in the circle c. N is the total number of alters (including duplicates if circles overlap) in all the circles.

We see that Group A achieves higher accordance and lower entropy than Group B. This suggests that the fine-grained circles in our exploratory visualization design are taken more holistically into consideration than Facebook smart lists by the participants to make visibility decisions. The larger difference in Entropy than in Accordance between the two groups is attributed to the fact that the participants perform particularly better with the large circles in Group A than in Group B. We also observe the performances decrease with increased α_{th} in both groups, which is understandable since the easy cases for visibility decisions are removed. Note that the performance of Group B decreases more than Group A. This indicates that the advantage of our visualization design is more prominent when the participants were making hard visibility decisions. The performance changes are summarized in Table II.

TABLE II. PERFORMANCE CHANGE WITH α_{th} RAISED FROM 1 TO 5.

	Group A	Group B
Decreased Accordance	0.054 (8.64%)	0.061 (12.22%)
Increased Entropy	0.024 (12.24%)	0.095 (20.61%)

V. CONCLUSIONS

A. Limitations and Future Work

Several limitations of the design were identified in the process of the experimental user study. First, some participants recommended to use photos instead of name labels of the alters. Second, the layout of the circles could be more compact

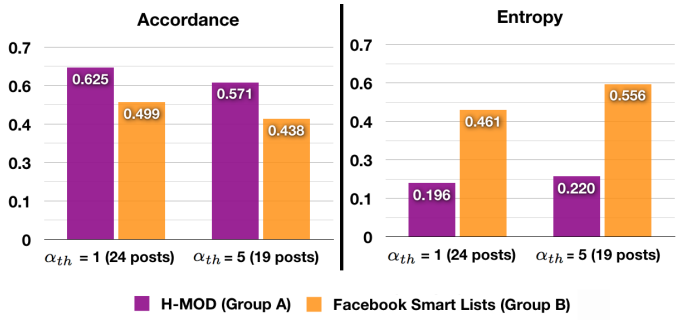


Fig. 5. Accordance and Entropy scores averaged over all posts in Group A and B, with $\alpha_{th} = 1$ and $\alpha_{th} = 5$. For a set of circles, the more it is in accord with the user's visibility decisions (Accordance) and the less bits of information needed to discern these decisions (Entropy), the better.

when the number of alters in a circle is small, so that the sub-circles in its parent would not overlap with other parent circles. Third, the participants, especially in Group A, were curious about the way that the circles were formed, which suggests us providing extra means to present the unique characteristics of the circles, such as labeling, showing the links among the alters, etc. Another limitation of this work is due to the limited number of participants in the user study. A larger sample size is needed for deeper statistical analysis.

B. Summary

A privacy concern in OSN is that users may be unable to well manage their online information flows due to a large number of contacts. In this paper, we introduce an exploratory application that leverages community discovery algorithm and visualization to help users make more effective decisions on the visibilities of their online posts. We describe an experimental user study to evaluate how effective is this approach to users. The positive results of the user study show that our approach is indeed useful in its regard.

ACKNOWLEDGMENT

The authors would like to thank their colleagues for providing their valuable comments and the large-hearted participants for contributing their private data and thoughts. The research presented in the paper has received funding from the Strategic Basic Research (SBO) Programme of the Flemish Agency for Innovation through Science and Technology (IWT) in the context of the SPION project (www.spion.me).

REFERENCES

[1] A. Acquisti, B. V. Alsenoy, E. Balsa, B. Berendt, D. Clarke, C. Diaz, B. Gao, S. Grses, A. Kuczerawy, J. Pierson, F. Piessens, R. Sayaf, T. Schellens, F. Stutzman, E. Vanderhoven, and R. D. Wolf, "D2.1- state of the art," Agentschap voor Innovatie door Wetenschap en Technologie IWT, Tech. Rep., Sep. 2011. [Online]. Available: https://www.cosic.esat.kuleuven.be/publications/article-2077.pdf

[2] K. Raynes-Goldie, "Aliases, creeping, and wall cleaning: Understanding privacy in the age of facebook." *First Monday*, vol. 15, no. 1, 2010.

[3] A. Lampinen, V. Lehtinen, A. Lehmuskallio, and S. Tamminen, "We're in it together: interpersonal management of disclosure in social network services," in *Proceedings of the 2011 annual conference on Human factors in computing systems*, ser. CHI '11. New York, NY, USA: ACM, May 2011, pp. 3217–3226.

[4] R. D. Wolf, R. Heyman, and J. Pierson, "Privacy by Design through social requirements analysis of social network sites from a user perspective," in *European Data Protection: Coming of Age*, S. Gutwirth, R. Leenes, and P. de Hert, Eds. Berlin, Germany: Springer, 2012.

[5] S. B. Barnes, "A privacy paradox: Social networking in the united states." *First Monday*, vol. 11, no. 9, 2006.

[6] d. m. boyd and E. Hargittai, "Facebook privacy settings: Who cares?" *First Monday*, vol. 15, no. 8, Aug. 2010.

[7] d. m. boyd, "Taken Out of Context: American Teen Sociality in Networked Publics," *Social Science Research Network Working Paper Series*, Feb. 2009.

[8] A. E. Marwick and d. m. boyd, "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience," *New Media & Society*, vol. 13, no. 1, pp. 114–133, Feb. 2011.

[9] B. Gao, B. Berendt, D. Clarke, R. D. Wolf, T. Peetz, J. Pierson, and R. Sayaf, "Interactive grouping of friends in osn: Towards online context management." in *ICDM Workshops*. IEEE Computer Society, 2012, pp. 555–562.

[10] M. Coscia, F. Giannotti, and D. Pedreschi, "A classification for community discovery methods in complex networks," *CoRR*, vol. abs/1206.3552, 2012.

[11] M. E. Newman, "Detecting community structure in networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 321–330, 2004.

[12] J. J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks." in *NIPS*, 2012, pp. 548–556.

[13] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial & Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[14] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[15] T. Schaffter and D. Marbach, "Jmod user manual," EPFL, Laboratory of Intelligent Systems, Tech. Rep., 2012.

[16] R. De Wolf and J. Pierson, "Researching social privacy on sns through developing and evaluating alternative privacy technologies," in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2012.

[17] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer, "Quantifying the invisible audience in social networks," in *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2013).*, 2013.

[18] J. C. Quiroz, S. J. Louis, A. Shankar, and S. M. Dascalu, "Interactive genetic algorithms for user interface design," in *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*. IEEE, 2007, pp. 1366–1373.

[19] Y. Rogers, H. Sharp, and J. Preece, *Interaction design: beyond human-computer interaction*. Wiley, 2011.

[20] S. Jones, "Automating group-based privacy control in social networks," Ph.D. dissertation, University of Bath, 2012.

[21] C. M. Brown, *Human-computer interface design guidelines*. Intellect Books, 1998.

[22] W. Lidwell, K. Holden, and J. Butler, *Universal principles of design: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Rockport Pub, 2010.

[23] M. Kurosu and K. Kashimura, "Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability," in *Conference companion on Human factors in computing systems*. ACM, 1995, pp. 292–293.

[24] K. Moore and J. C. McElroy, "The influence of personality on facebook usage, wall postings, and regret," *Computers in Human Behavior*, vol. 28, no. 1, pp. 267–274, 2012.

[25] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor, "I regretted the minute i pressed share: A qualitative study of regrets on facebook," in *Proceedings of the Seventh Symposium on Usable Privacy and Security*. ACM, 2011, p. 10.

4

PAPER 2:

Friends and Circles – A Design Study for Contact Management in Egocentric Online Social Networks

Friends and Circles — A Design Study for Contact Management in Egocentric Online Social Networks

Bo Gao and Bettina Berendt

Abstract Users in Egocentric Online Social Networks (EOSN) may share private information with the "wrong" friends. To mitigate this problem, we first designed an exploratory visualization for friend-grouping. We then conducted a user study, through which we found that, comparing Facebook smart lists, the hierarchal modularity-based communities were more helpful for users to make visibility decisions in online posting. We then compared the modularity-based algorithm (MOD) with another state-of-the-art community detection algorithm. The results showed that the ground-truth circles coincided more with the MOD-circles. We further extended MOD to produce overlapping circles and found even better results. Furthermore, informed by our user study, the research on social groups and information visualization theories in general, we developed a friend-exploration/grouping web application for Facebook users.

1 Introduction

An Online Social Network (OSN) today can hold hundreds of millions of users. Two years ago (2012), Facebook (www.facebook.com) has reached its "one billion users" mark [64]. Behavioural and sometimes very personal information of OSN users is uploaded and shared online daily, in large quantity and tremendous detail. While the availability of these data enables us to understand more about our societies, it also challenges us in effectively and efficiently processing large amount of information, and managing our online personal content.

Bo Gao
Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium.
e-mail: bo.gao@cs.kuleuven.be

Bettina Berendt
Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium.
e-mail: bettina.berendt@cs.kuleuven.be

In the age of online social networking, "even as bloggers and networkers delve into their private experience, they communicate with their fellow humans in a shared festival of the self" [5]. Such phenomenon has raised concerns about privacy. For example, it was found that OSN users had demonstrated high privacy concerns while revealing great amounts of personal information [1]. It also has been quantitatively demonstrated that users' perceptions of audience size do not match reality, since not enough feedback is provided [8]. boyd showed that collapsed or ambiguous online contexts could lead to undesired disclosure of personal information [10]. Gürses extends the notion of privacy from confidentiality to access control and practice. The extended notion encompasses the solution space in which OSN users are empowered to re-negotiate the boundaries of information dissemination and construct their online identity based on a transparent system [30] .

In light of the recent privacy research, we become interested in the tools that can help OSN users gain insights into their own social networks, explore to reveal hidden patterns and control the flow of personal data shared with online friends. As previous studies have suggested [37, 48, 26], in order to manage personal information flow, it is important for users to categorize their online friends into groups, categories, circles, lists or communities[1], so that the user can post towards clearly specified audience. By "post", we mean the user's action of uploading or sharing digital information in OSN. We will also be using the term Egocentric Online Social Network (EOSN) to refer to a sub-network in an OSN, with the nodes representing people and the (directed or undirected) edges representing certain relationships among them. The network is centered on one user (as the ego), whose friends (as the alters) are directly linked to this user via edges. Edges usually also form among the friends.

As reported in 2011, the median number of friends of a Facebook user was 100 [57]. This number became 229 in 2013. For teens and people in their 20s, it was 400 or more [62]. To make sense of the increasingly complex online social networking data and manage online contacts, a user needs to deploy more sophisticated tactics (categorizing friends under different situations) than simple browsing and memorising. We started looking into visualization approaches to address this issue, as human visual system is highly parallel and pre-attentively sensitive to variations in visual stimuli, such as color, shape, positions, etc. [52]. With a carefully designed interactive visualization system, the user should be able to gain an overview of her network, explore the network to find novel patterns and easily construct groups of friends for different posting purposes.

The contributions of this chapter are: First, we documents a user study, which shows that, compared with Facebook smart lists, the hierarchical modularity-based circles (used in an exploratory fashion) are more supportive for users to make privacy-related visibility decisions in online posting. Second, we design a new form

[1] We use these words interchangeably throughout the paper. The words "group" and "category" are used more generically, "list" is often used in the context of Facebook and Twitter (www.twitter.com). We use "circle" more often in the context of Google+ (www.plus.google.com) and visualization. The word "community" is usually used in the context of community detection algorithms.

of interactive visualization to visualize hierarchically grouped items. The items can be individual friends a user has in her online social network. Third, we test two community detection algorithms on three egocentric social network datasets. The results show that the ground-truth circles coincide more with the modularity-based circles. Fourth, we extend the modularity-based algorithm to accommodate the overlapping nature of online social circles. The experimental results show that this approach is indeed better than the original modularity-based algorithm. Fifth, we develop a friend-exploration/grouping web application for Facebook users to explore their online social networks and create their customized friend-lists.

The structure of this chapter is as follows: Section 2 covers a set of existing tools for EOSN analysis and friend grouping. In Section 3, we analyze users' requirements, motivate and detail a new design of interactive visualization, named CircleTree. We then use it as the common interface to conduct a user study. The study compares users' behaviour in privacy decision-making based on two different friend-grouping mechanisms. In Section 4, we examine two community detection algorithms and discuss the nature of friend grouping in EOSN. We then propose an extended version of the modularity-based community detection algorithm to generate overlapping friend groups. In Section 5, with the introduction of the tool named FreeBu, we propose alternative views to supplement the earlier CircleTree visualization. We then identify the improvement points for the friend-exploration/grouping tool design. In Section 6, we conclude with a summary and an outlook on future work.

2 Related Work

We are interested in the tools that enable users to gain insights into their EOSN and/or construct friend groups. We describe a selective set of existing tools in Section 2.1, and discuss their relationships with our contributions in Section 2.2.

2.1 Existing Tools

PViz [38] is a tool that helps Facebook users understand their privacy settings. The tool is compared with existing policy comprehension tools on Facebook, namely Audience View and Custom Settings. It was shown that PViz was more effective for the users in comprehending privacy settings. Privacy Wizard [19] is a tool that can automatically predict the privacy preferences for a Facebook user based on her previous privacy-setting input, the result is also encouraging. Both tools employ Newman's Modularity-based community detection algorithm [46] to drive friend groups, which are then used for visualization (PViz) and prediction (Privacy Wizard) respectively.

NodeXL [53] is a general-purpose plugin that allows users to draw graphs by using a Microsoft Excel template. It implements various graph clustering algorithms, including modularity-based ones. It supports social network analysis, users can visualize their Facebook and Twitter graph data via an importer interface. The tool uses the Group-In-a-Box (GIB) feature [49] to help users delineate the clustering structure of the imported graphs. More specifically, the visual clusters are firstly formed in a graph layout. They are further constrained by being placed inside boxes whose sizes depend on the respective numbers of nodes. These boxes are then arranged by the squarified treemap algorithm [11]. The GIB layout is also used for multivariable grouping of the nodes based on their attributes. Some other general-purpose network-analysis tools are potentially useful for OSN users as well, such as Gephi [6], Cytoscape [51] and Tulip [3].

There also exist many small web applications that visualize OSN users' network data and allow simple interactions for exploration. Here we give two representative examples. Social Graph[2] is a Facebook application that shows a force-directed layout of the user's friend graph. The nodes are colored according to the detected communities based on Modularity [46]. The user can click on an individual friend (i.e. a node) to see the friend's profile photo along with three statistical numbers: the number of mutual friends she shares with that friend, the clustering coefficient of the friend and the clustering coefficient of the corresponding community. The user can further explore the graph layout by selecting a specific community from a drop-down list, so that only the members from that community are repositioned and displayed on the screen. Each community is labeled with the name of the friend in that community who has the highest clustering coefficient. The other example is InMaps[3]. It is a web application similar to Social Graph. It visualizes the user's network on LinkedIn (www.linkedin.com) with rather similar force-directed layout and modularity-based communities as well. Through InMaps, a LinkedIn user can zoom and pan to explore the map. The name labels of the friends are simultaneously brought to display upon zooming-in. We can also see that the nodes and labels are mapped with care to avoid overlapping, which makes the visualization more readable than Social Graph.

Personal Analytics for Facebook[4], as part of the Wolfram Alpha knowledge Engine (www.wolframalpha.com), is a state-of-the-art visual and textual analytic web application designed for Facebook users. It offers a wide range of analytics, including various friends' demographic reports, summaries of the user's logging-in, posting and sharing activities, etc. Another merit of this tool is that each analytic segment can be downloaded in different formats for other uses, such as spread sheet, image and vector graph.

Furthermore, current Social Networking Sites (SNS) provide mechanisms for users to create their own friend groups, such as the lists in Facebook and Twitter, the circles in Google+ and the groups in Weibo (www.weibo.com). But by large, users

[2] https://apps.facebook.com/socialgraph_fr_yl [Accessed on Nov 30, 2013]

[3] http://inmaps.linkedinlabs.com/network [Accessed on Nov 30, 2013]

[4] http://www.wolframalpha.com/input/?i=facebook+report [Accessed on Nov 30, 2013]

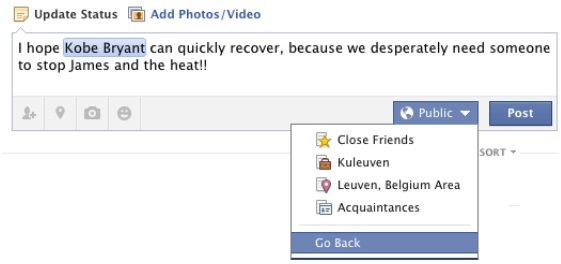


Fig. 1 A Facebook user can conveniently limit the visibility of her status by choosing one of the four lists, of which *Close Friends* and *Acquaintances* are the lists that the user manually defines, and *Kuleuven* and *Leuven, Belgium Area* are the automatically generated smart lists, based on the user's work and current city.

have to manually group friends, which tends to become unmanageable. We know one exception – Facebook "smart lists" — that can automatically generate friend lists. Facebook smart lists[5] provide users with an automatic grouping solution. The lists are generated based on the information about the user's education, work and current city. For example, if the user indicates Leuven as her current city, she will have a list with all of her friends who also indicate Leuven as their current city. The user can directly determine the audience of her posts by choosing one of the lists, including the smart lists. Figure 1 gives an example for status update.

2.2 The Tools' Relations to Our Contributions

Informed by PViz and Privacy Wizard, we find that one feature of Facebook's privacy control mechanism yet to be examined is the smart lists. In Section 3, we take the smart lists as baseline and investigate the roles that community detection algorithms and interactive visualization play in users' privacy decision-making process. We also note that PViz is for privacy-setting comprehension, Privacy Wizard is for privacy-setting configuration, both tools do not serve for the purpose of helping users create their own friend groups, e.g. Facebook friend lists. We built the friend-exploration/grouping tool (as detailed in Section 5) to facilitate this activity.

The GIB feature in NodeXL currently does not support hierarchical graph clustering and exploration. A hierarchy is difficult to visualize and interact with, because the semantics from different layers may compromise the readability of a set of visual clusters, especially when the leaf nodes are of main interest (e.g. the user' friends). In Section 3, we introduce a hierarchical exploratory visualization design. This design displays the grouping structure of the user's EOSN and maintains the emphasis on the leaf nodes of a hierarchy.

[5] https://www.facebook.com/help/204604196335128/ [Accessed on Nov 30, 2013]

Unlike Gephi, Cytoscape and Tulip, NodeXL offers its users connectivity to their Facebook accounts, so that they can easily analyze their ego-networks. Besides such connectivity, our tool also provides its users with a series of list-creation interfaces. The user-created friend lists can be submitted to their Facebook accounts. Moreover, NodeXL, Gephi, Cytoscape and Tulip are desktop applications/plugins that require installation. For NodeXL, the installation is conditioned on having Microsoft Excel in advance. Our tool is an online application that is easily accessible by a JavaScript-enabled browser. To facilitate user-defined friend-grouping, in Section 5, we elaborate the ways in which we improve the existing graph-layout visualizations, such as Social Graph and InMaps.

In Personal Analytics for Facebook, we find that the visualizations are fairly static as it follows the interaction syntax of a regular web page – that supports up/down scrolling and hyperlink clicking, but without zooming, panning and animation. Thus the action of inspecting individual objects in an overview visualization, e.g. foraging through the graph clusters, become problematic if the user has many friends. We consider our interactive visualization design to be complementary with respect to this.

Furthermore, we notice that all the aforementioned tools in Section 2.1 use or include modularity-based communities to approximate the user's social groups. Modularity maximization encourages mutually connected nodes to be put into the same community. While the broad adoption of this method is partly due to its popularity and software availability, another contributing factor seems to be that, among many other community detection methods, it produces the communities that best match the communities a user has in mind. Various studies have demonstrated useful applications of modularity-based community detection algorithms for social network analysis [46, 35]. We will also be using this method in our user study (Section 3). In Section 4, we examine this method in more detail and demonstrate its usefulness for EOSN friend-grouping. We also propose an extension of the modularity-based algorithm and show that it has a better performance.

3 A User Study on Circles for Visibility Decisions

This section consists of three parts: First, in relation to our user study, we discuss OSN users' need for friend-grouping tools (Section 3.1) and why we further choose to use hierarchical grouping (Section 3.2). Second, we examine the related work on visualizations for hierarchies (Section 3.3) and describe the CircleTree visualization that we have developed (Section 3.4). This visualization is then used in our user study. Third, we describe the participants, tasks and results of the user study (Section 3.5).

3.1 The Need for Grouping Tools

As discussed in Section 1 and 2, we know that the increasingly large amount of data produced by our online social networking activities has made it difficult for us to manage our personal information flow. Sharing certain information with the wrong people can cause awkwardness, embarrassment or even severe damage on the user. Therefore a tool is needed to inform OSN users and facilitate their privacy decision-making. More specifically, the user should be able to effectively determine which piece of her personal information is visible to which friend(s). But it would be a daunting task if the user goes through each individual online friend that she has one by one, and considers that friend's unique constellations of attributes and proclivities in order to make such a decision. In reality, informed by the research in social cognition [36], we know that people "prefer to construe others on the basis of the social categories to which they belong, categories for which a wealth of related material is believed to reside in long-term memory". Because of the limitations in human cognition and the challenges presented by a vast stimulus world (in our case – the online social networking environment, intertwined with the offline social life), a person naturally employs categorical thinking in order to simplify and structure the people she befriends [2, 36]. This description further provides support for the necessity of friend grouping [37, 48, 26].

We will be using the term Visibility Decision to refer to a user's binary decision on whether a post is visible to an individual friend in her EOSN. A post can be anything that a user uploads or shares in an OSN, e.g. a status update, a (re)tweet, a photo, a comment or an article shared, etc. Friend grouping can facilitate users' visibility decisions. The user decides the visibility of a post directly based on friend groups rather than individuals. In other words, when the user sees a group, assuming her previous familiarity with the group, she can skip the serial browsing that examines individual friends in this group, and determine the visibilities of the post towards those friends on a group level. There are two exceptions in which the user does not directly deploy the groups in her visibility decisions. *First*, certain posts are too privacy sensitive, i.e. it has become a complete regret, or not sensitive at all, in both cases, a binary decision becomes unary, and all user's friends are considered as one group. *Second*, when the number of friends who are or are not supposed to see a post (e.g. one, two or three) is significantly smaller than the number of friend groups shown to the user, then checking the groups requires more effort than just doing a standard search, e.g. typing friend names in a search box. Thus it is no longer necessary to use groups. However, we shouldn't completely disregard friend grouping in such situation, because it can raise the user's awareness about her friends, which can be useful for other aspects of life or later visibility-decision-making. Moreover, if shown appropriately, the friend grouping can help the user spot "unexpected" or "surprising" friends, which then becomes useful for the user to make visibility decisions.

3.2 Why We Use Hierarchical Grouping

In our user study (Section 3.5), we compare the two ways of detecting communities for a Facebook user – Facebook Smart Lists (FSL) and hierarchical modularity-based communities – in terms of their usefulness in facilitating the user's visibility decisions for posting. The original modularity-based community detection algorithm (MOD) takes the user's friend graph as input and produces non-overlapping, flat communities. There is a subgraph corresponding to each detected community of nodes. MOD is then applied to each subgraph, deriving sub-communities. We adapt MOD into a hierarchical one, abbreviated as HMOD. We choose to use HMOD for three reasons: (1) Modularity-based methods are known to have a "resolution limit" problem [22]. It is most likely that, for a community with $\sqrt{m}$ (m is the total number of edges) or less nodes, its sub-communities cannot be discovered. This implies that modularity optimization can miss the substructures of a network. (2) It is well known that people organize semantic concepts hierarchically in memory [13]. The reason for this is because storing generalized information with superset nodes is more economical for humans. Hierarchy is necessary in the navigation for the retrieval of more detailed information. (3) Another incentive that we use HMOD is based on the aforementioned Categorical Thinking, as iterative categorization (i.e. grouping) may be required from the user to make sense of the her friends if the number of friends is simply very large.

3.3 Related Work on Visualizations for Hierarchies

To examine the difference between two friend grouping strategies, there needs to be one common User Interface (UI). Given that a Facebook user usually has hundreds of friends, naively using "pen and paper" to elicit the visibility decisions from the participants may weary them. Bearing this in mind, we decide to let the participants operate on a computer-based UI. For users making visibility decisions with such an interface, we need two basic functions: *First*, browsing is applicable at both group and individual levels. *Second*, making a decision is applicable at both group and individual levels. Various existing works have paved the way for visualizing hierarchical grouping structure. We do not intend to provide a comprehensive review in this subsection. Instead, we give a qualitative treatment to four representative types of visualizations and motivate our design choices. We refer to the two dimensional area on the computer screen where a visualization is rendered as the canvas.

- **Node-Link Diagram** The traditional Node-link Diagrams use shapes (rectangles, circles, etc.) to represent nodes and lines to represent links. The direction from the root of the tree to the leaves is either vertical or horizontal. The nodes (intermediate or the leaves) at the same level need to be aligned at the same vertical or horizontal line. Hence only one-dimensional space is utilized to visualize each level. As shown in Figure 2a, this space can be easily exhausted, especially

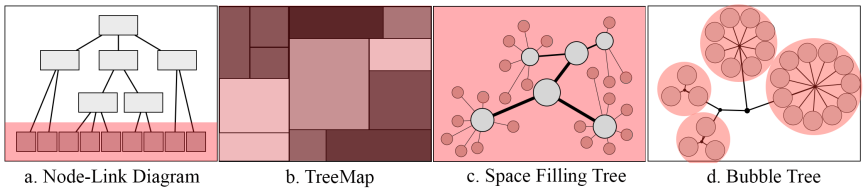


Fig. 2 Four types of representations for visualizing a hierarchical grouping structure, the potential area that can be used to draw leaf nodes is overlaid with red color.

at the leaf level. When the leaves are squeezed to be aligned and fit into the canvas, they easily become too small for the user to interact with, and the grouping structure is no longer clear at the leaf level. Improvements have been made using coloring and merging to reduce the number of branches and/or leaves to draw (e.g. Colored trees [50]). However, they leverage the continuous values of leaves, so that the colors correspond to different average values, giving a sense of numerical ordering. In our case, either the friends or the groups are discrete, which the user needs to differentiate to make a visibility decision. We also note that using the color visual channel to differentiate discrete variables (e.g. Stacked Tree [9]) is problematic, as there are very limited choices for visually distinct colors [31, 28].

- **TreeMap** Grid-based (or matrix-based) visualizations utilize the canvas space more efficiently, as shown in Figure 2b. A typical grid-based layout is treemap [33]. It visualizes hierarchical data by nested rectangles. Many techniques have been proposed to make treemaps more structurally perceivable by humans. For example, shaded colors can bring a sense of ordering to the treemap nodes [54], gradient colors can demarcate different clusters in a treemap (cushion treemap) [58], the aspect ratio of the nodes can be adjusted to improve their readability (squarified treemap) [11]. Compared with node-link diagrams, treemap is more readable for various large-graph-related tasks, but path finding is consistently in favor of node-link diagrams [27]. More importantly, the user cannot conveniently select all the friends in a (sub-)group at once to make a visibility decision in treemaps.
- **Space-Filling Tree** Given the limitations in node-link diagrams and treemaps, hybrid visualizations have been proposed. The space-filling tree [47] is a typical example, as shown in Figure 2c. It spreads the nodes and leaves across the whole canvas. To give a sense of structure, the sizes of the nodes decrease with ascending levels of the tree, the child nodes are mapped in proximity with their parent. But it is probable that, in order to optimally utilize the unoccupied space, the nodes in one branch protrude into the neighborhood of another branch, resulting in a less structural display.
- **Bubble Tree** To heighten the sense of grouping structure, Bubble Tree [29] further constrains the proximity mapping between child and parent nodes – the child nodes are aligned in a circle around their parent, as shown in Figure 2d. This sacrifices potential drawing area on the canvas (still more space-filling than the tra-

ditional node-link diagram), but gains the representation of a stronger grouping structure. The user can select a branch of nodes via their parent node. However, it remains difficult to compare the sizes of the branches on the same level.

3.4 The CircleTree Exploratory Visualization

Considering the previously examined visualizations, we realize that showing the complete structure of a tree of friends may be unnecessary, even interferential to the user. As we try to facilitate the user in determining whether a friend (represented by a leaf node) can see her post, drawing too many intermediate nodes on the canvas produces unnecessary "cognitive overhead" [7], because those nodes not only occupy limited canvas space, but also increase the number of objects that the user needs to process in the limited short-term memory. Therefore, we design a new form of interactive visualization that constrains the number of levels shown (namely one or two levels) and let the user's zooming actions reveal more sub-groups or less only when she needs to. It is also similar to the Bubble Tree in the way that child nodes are positioned in a circle around their parent. We call it CircleTree.

It is important to note that in order to make visibility decisions for a post at the very beginning, the user needs to go through all the friends, regardless of the form of presentation, either simple textual list on a paper or complex visualizations. The benefit of a (good) friend grouping follows after the user's initial contact and familiarization with the generated groups. In other words, the user has made the connection between members and their corresponding group. A group is represented by a token, which can be a shape, a descriptive phrase, or the name of a member from this group, etc. This linkage information is stored in the user's long-term memory. The members can be recalled when the user just sees the group token. In such a way, the user bypasses the serial browsing of each individual member, and directly utilizes a group. The main purpose of our visualization design – CircleTree – is to provide visual tokens for a user's friend grouping. It also adds the elements of structure and engagement to an otherwise lengthy, textual reading and decision-making task. Another purpose of the visualization is to facilitate manual friend-grouping construction, as elaborated in Section 5. The CircleTree visualization is detailed as follows:

A node is represented by a circle, a group of nodes is represented by the circular placement of its child nodes around one extra node, which is the parent node that represents the whole circle, as shown in Figure 3a. With the basic visual principles in mind – that humans are very sensitive to the difference of lightness in grey colors [55], we set the background color white, the friend nodes grey, the parent nodes blue. The latter two colors are also semi-transparent to avoid the occlusion effect. We pick orange and magenta as the highlight color for each friend node and parent node respectively. The large differences (from grey) in saturation and (from blue) in hue promote visual contrast [59]. At first sight, it seems sufficient to use just one visual channel to encode grouping, i.e. the circular placement of child nodes in a group.

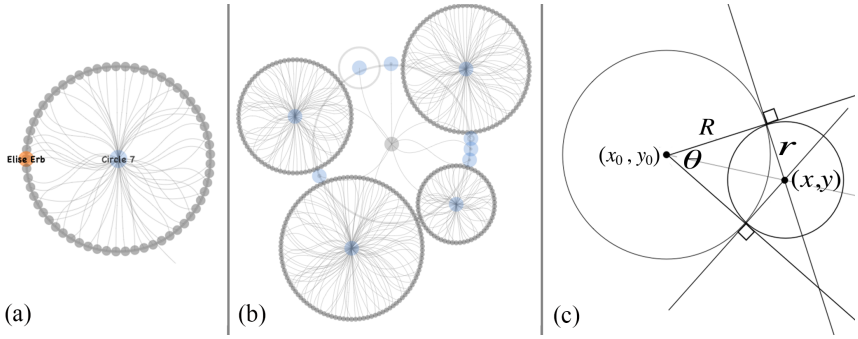


Fig. 3 (a) A single group circle, the grey nodes are the friend nodes, the blue node is the parent of the group circle. (b) The groups are positioned around a central node. (c) An illustration of drawing a group circle around a central node.

But since the user is allowed to drag the nodes to other positions on the canvas, as described below, we add lines connecting the child nodes with the corresponding parent to emphasize that a child node belongs to its parent. The lines within a circle also signal a sense of integration. But in order to avoid overemphasizing the lines instead of the nodes, and sometimes to avoid occlusion between lines and nodes, we choose to increase the transparency of the lines[6]. Furthermore, as argued, curved lines can be used to make certain paths in a graph more apparent [61], based on [20], and curved shapes are often reflective of natural objects, giving the observer a pleasant feeling [34], we choose to use Bézier curves instead of straight lines. However, the exact role that curves play in improving the perception of grouping structure and the aesthetics of the visualization is unclear, and beyond the scope of this work.

The groups are then positioned approximately in a circle around the root node that is under focus, as shown in Figure 3b. In the initial layout, this top node is the root of the tree. We see that the circumference of each group circle formed by its child nodes is naturally scaled with the number of children, presenting a visual order. Every pair of adjacent group circles are tangent to each other. The CircleTree layout algorithm is detailed in Algorithm 1. The radius r_i of an individual friend node from a group circle c is then approximated by $r_i \approx \pi \cdot r/|c|$, where $|c|$ is the number of friends in c, r is the radius of c. Note that very large or small m results in an exceptionally small or large r_i. Thus, minimum and maximum radii r_{min} and r_{max} are set to prevent each friend node from being too small to see or too large that it disturbs the visual ordering. When c has few friends, its assigned r becomes small, making $r_i < r_{min}$. After restoring the overly small r_i to r_{min}, we will likely have relatively large child nodes occupying the entire inner space of c and overlapping with the central parent, which does not make sense to show. Therefore such friend nodes are automatically hidden from sight, instead, the user will only see the grey

[6] Note that this intended reduction of opacity does not make the lines difficult to see on a computer screen, but may lead to sub-optimal printing quality.

Algorithm 1 The algorithm for computing the layout of the group circles around a center (x_0, y_0). Note that (x_0, y_0) can be the position of the root or any center of a parent node of a group circle. We also set the maximum angle for each circle to $\pi/2$, which is an empirically derived value to keep the sizes of the generated circles contained within the canvas. For symbols θ, x_0, y_0, x, y, r and R, please refer to the illustration in Figure 3c.

Require: the array Arr storing the sizes of the circles.
n = No.Circles, N = No.Friends, $MaxAngle = \pi/2$.
Let the array $Angles$ store the angles θ the circles.
for $i = 0$ to $n-1$ **do**
    $Angles[i] = 2\pi \cdot (Arr[i]/N)$
    if $Angles[i] > MaxAngle$ **then**
        $Angles[i] = MaxAngle$
    end if
end for
Let the array CS store the tuples (x, y, r).
if n > 1 **then**
    $x = x_0 + |\tan(Angles[0]/2) \cdot R|$
    $y = y_0 - R$, $r = x - x_0$
    $CS[0] = (x, y, r)$
    $totalAngle = Angles[0]$
    for $i = 1$ to $n-1$ **do**
        $r = |\tan(Angles[i]/2) \cdot R|$
        $s = \sqrt{(r^2 + R^2)}$
        $x = x_0 + \sin(totalAngle + Angles[i]/2) \cdot s$
        $y = y_0 - \cos(totalAngle + Angles[i]/2) \cdot s$
        $CS[i] = (x, y, r)$
        $totalAngle = totalAngle + Angles[i]$
    end for
else
    $CS[0] = (x_0, y_0, R)$
end if
return CS

circular silhouette around the parent to mark the visual area of the group, keeping the visualization clean and ordered, as illustrated in Figure 3b.

In the visualization, initially, the user only sees one layer of the tree, as an overview, but can further explore it by the zooming, panning and enabling text labels. We assume that a user can recall her impression of or her relationship with a friend if she see that friend's name. Therefore, when the mouse hovers over a node, the node is highlighted and corresponding label is shown, either a friend name or the name of a numbered intermediate node (e.g. "Circle 5" or "Circle 5.3"). Right-clicking on a parent node maps its child nodes (which we call "the focused children") in a grid layout with the names brought into sight. When a grid layout is triggered, we increase the transparency of all the other nodes on the canvas, so as to reduce the interference from irrelevant visual objects, but still keep them visible in the background to maintain a global context, as shown in Figure 4a. Clicking (left or right) anywhere other than "the focused children" or another parent node

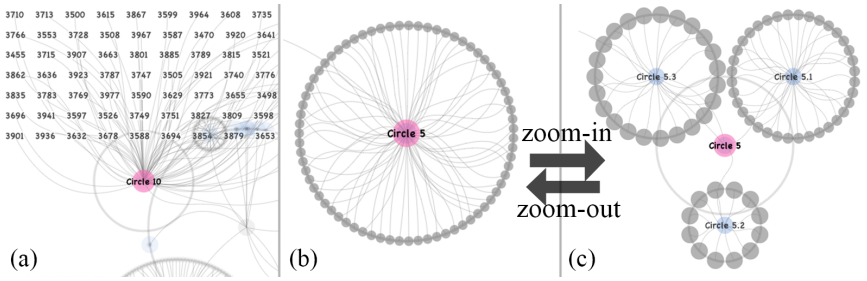


Fig. 4 (a) Right-clicking a parent node reveals the names of friends in that group. (b) A group of friends before zooming-in. (c) The same group of friends from (b) who are further grouped after zooming-in.

on the canvas will restore the original layout. Right-clicking on another parent node will automatically restore the circular placement of the currently focused children, meanwhile shift focus onto the children of the newly clicked parent node. The user can pan (drag to displace visual objects) to adjust the point of interest. If the starting point of panning is not over a node, the whole tree will be panned. If it is over a node, that node will be panned, along with its child nodes if it is a parent.

We take the current mouse position on the canvas as the "anchor point" for zooming actions. An anchor point $P_{anchor} = (x_a, y_a)$ is the position that is invariant during zooming. A zooming action triggers the following transformation: $rt \cdot \beta \cdot (P' - P_{anchor}) = (P' - P)$, in which $P = (x, y)$ is the position before zooming, $P' = (x', y')$ is the position after zooming, $rt \in \mathbb{R}$ is the value of mouse-wheel rotation provided by the operating system, $\beta \in \mathbb{R}$ is a constant adjusting the zooming speed. Note that the zooming speed on X- and Y-axes are the same. It then follows that the scaling factor is $sf = (x' - x_a)/(x - x_a) = (1 - rt \cdot \beta)^{-1}$. During zooming, the radius r_i of each node is multiplied by sf but further constrained by $r_i \in [r_{min}, r_{max}]$. When the child nodes no longer overlap with the corresponding parents, the hidden child nodes and their names are brought into display with zooming-in. When the user zooms into one circle of friends, we perform a "focus-check" to determine whether to further divide the circle. The "focus-check" assumes a rectangular area, half the width and height of the canvas, with the current mouse position as the center point. Upon the user's zooming-in, the only remaining group circle whose parent node is within this area is found and divided. The newly generated sub-circles are presented if the subgraph corresponding to the circle is divisible according the algorithm [46]. We choose the size of the "focus-check" area such that the user does not need to zoom too deeply or too shallowly to explore sub-circles. Zooming out of the visualization makes the sub-circles from the previously divided circle squeezed and overlapped, which will trigger them to merge back to the singular circle again. This is depicted in Figure 4b and 4c.

3.5 Participants, Tasks and Results

In this section, we document the tasks of the participants and the findings from the study.

3.5.1 Participants and Tasks

There were 16 participants (three females), 25-45 years old, from eight countries. Among them were Ph.D students, company employees and graduate students. The participants were equally divided into two groups, which we named directly with the corresponding algorithm abbreviations mentioned in Section 3.2 – HMOD and FSL. Both groups used the same visualization interface detailed in Section 3.4, but with different community detection methods, as their names suggested, namely the hierarchical modularity-based algorithm and Facebook smart lists. Because the latter was not a complete grouping, the friends of a participant that were not in any smart list were put together as one other group.

Our assumption in the user study is that users utilize categories of friends (denoted as C_u) to make a binary visibility decision. We denote the communities that HMOD and FSL produce as C_{HMOD} and C_{FSL} respectively. C_{HMOD} is the result of the interactions between a user and HMOD, with the CircleTree visualization interface. C_{FSL} is the set of non-hierarchical circles of friends constructed from the user's Facebook smart lists, with one extra circle containing the friends who are not in any of the smart lists. Our hypothesis is that, for users' visibility decision-making, C_{HMOD} coincide with C_u, more than C_{FSL}.

We asked the participants to perform the following two tasks: elicitation of regrets in posts and visibility decision-making. In the first task, each participant was asked to identify her regretted posts. In the second task, the participants in the two groups HMOD and FSL were asked to make visibility decisions for each of their posts. Each group has 8 participants and 24 posts. As illustrated in Figure 5, when a participant thinks a friend can see the post, she clicks on the corresponding friend node, the color of which changes to indicate that the post is now visible to the clicked friend. Clicking on the parent node of a group circle toggles every child node's color, or further descendants if some child nodes are already divided by a zooming-in action. The participants could work at their own pace until they were satisfied with their decisions.

Though recent studies have investigated regrets in OSN from different aspects [41, 60], we chose to let the participants explicate their own regrets, as it is easier for a person to make visibility decisions based on her own experience. We collected the posts in face-to-face interviews with the participants. We emphasized the difference between complete and partial regrets. A complete regret meant that the post was supposed to be seen by no one. A partial regret meant that the participant did not mind her post being seen or intended her posts to be seen by some of the friends, but failed to block the other undesired friends. Since a complete regret entails concealing the corresponding post completely, which would render a visibility decision

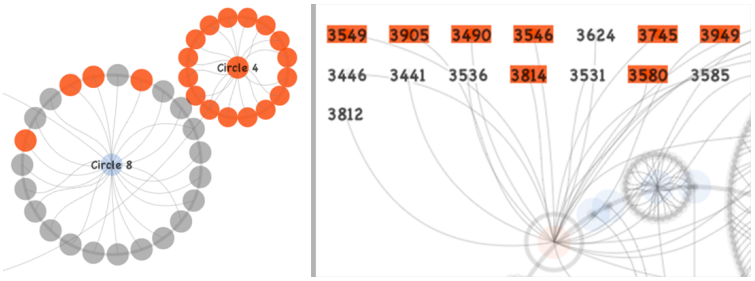


Fig. 5 Participants can determine a post's visibility to each friend individually by clicking friend nodes or collectively by clicking parent nodes in the centers of group circles.

trivial, we guided the participants to only think of partial regrets. Each participant was encouraged to think of at least three posts. A post needs to be specific enough to let the participant define its visibility towards each friend. In total, 48 posts were collected; each participant contributed three personal posts on average. We found that photo-related posts were mentioned frequently, thus making a distinction between photos and topics. Topic-related posts include status updates, web-link sharing and comments.

We recorded the participants' regretted posts and manually classified them into five categories, as summarized in Table 1. The *first* category covers the posted photos that cause embarrassment or awkwardness, typical examples are "drunk party" photos. There are also the photos showing the participant together with some particular person(s), e.g. ex-boy/girl-friend, that the participant feels the need to hide the photos from some friends. The *second* category covers the photos that are less sensitive in terms of embarrassment or awkwardness, but still in need of visibility control. For example, some photos may be so intimate that the participant only wants to show them to her family and best friends. Some photos were taken at a event with a specific group of people, only to whom, as participants argues, the photos should be made visible. More than a third of the posts are photo-related. The *third* category covers the topic-related posts that involve explicit self-expression, including strong opinions and emotional expressions, such as venting negative emotions. Of the seven posts in this category, six are about venting or expressing negative opinions, which the participants felt should be avoided in future, for those posts may harm one's image if disclosed carelessly. The *fourth* category covers the sensitive topics that are less self-involved, but more about the intrinsic sensitive nature of the content of the posts, including politics, religion, sex, race and/or nasty jokes. It is interesting to see that nine out of the twelve posts in this category are about inappropriate jokes. For example, several participants reported that they posted something they believed sarcastically humorous, but in hindsight, they thought it was not wise to expose those posts publicly, as some friends may not understand the humour, or even be offended by it. The *fifth* category covers the relatively less sensitive topic-related posts, which nonetheless need visibility control. For instance, it may not

Table 1 Participants' Regretted Posts

	Categories of Regretted Posts	Frequency
(1)	sensitive photos causing embarrassment or awkwardness	8
(2)	other photos for a specific group of friends	9
(3)	sensitive topics involving emotional expressions	7
(4)	sensitive topics involving nasty jokes	12
(5)	other topics for various specific situations	12

make sense to show the posts to the friends who do not speak the language in which the posts are written.

3.5.2 User Study Results

We use binary entropy to evaluate the effectiveness of the two approaches for users making visibility decisions. $Entropy(post) \in [0,1]$ (Equation 1) calculates the information content (in bits) needed to determine whether a member in a circle can see a post. C is a set of circles of friends, and $c \in C$ generated by HMOD or FSL. $V_{c,post}$ is the number of the friends to whom $post$ is visible in the circle c. N is the total number of friends (including duplicates if circles overlap) in all the circles. $Entropy(post) = 1$ means that on average, in one circle, half the circle can see the post while the other half cannot. This indicates that the given set of circles is unhelpful for the user to make visibility decisions on a group-level, by taking the circles holistically into account. $Entropy(post) = 0$ means that for each circle, the friends in the same circle have the same visibility access to the given post. That is, every circle can be fully utilized by the user to make visibility decisions. The CircleTree visualization in the group HMOD is analogous to a binary-classification tree. Users try to use this tree to make visibility decisions. A"pure" circle in terms of visibility decisions is helpful, since such a circle can be considered as a whole. The initial circles are divided until they are indivisible according to the graph modularity or they are pure. Then the sub-circles are used to calculate entropy scores.

$$\begin{aligned} &Entropy(post) = \sum_{c \in C} \frac{|c|}{N} Entropy(c, post) \text{ with} \\ &Entropy(c, post) = -\frac{V_{c,post}}{|c|} \cdot log_2 \frac{V_{c,post}}{|c|} - \frac{|c| - V_{c,post}}{|c|} \cdot log_2 \frac{|c| - V_{c,post}}{|c|} \end{aligned} \tag{1}$$

Another aspect of a set of visibility decisions for a user's post is its imbalance. That is, the number of friends who can see the post is significantly different than those who cannot see the post. Let V_{post} be the total number of friends who can see the post and $\alpha = min(V_{post}, N - V_{post})$. When α is rather small, e.g. one or two, $Entropy(p)$ can be low almost regardless of which grouping method is used. In such

Table 2 Entropy scores for group FSL and group HMOD, with $\alpha > 1$ and $\alpha > 5$.

	FSL	HMOD
$\alpha > 1$ (24 posts)	**0.46**	**0.20**
$\alpha > 5$ (19 posts)	0.56	0.22

case, while a grouping may still be useful for the participants to browse friends, but it is likely to be less effective for making visibility decisions than the participants just typing individual friend names to search for them in real-time, as discussed in Section 3.1. We know that the average number of friends of each participant is 194. All the 48 posts (24 posts for each group) have $\alpha > 1$ and $\overline{\alpha} \approx 34$. Within these posts, there are 38 posts (19 posts for each group) with $\alpha > 5$ and $\overline{\alpha} \approx 42$. Table 2 shows the average Entropy scores in group HMOD and FSL for $\alpha > 1$ and $\alpha > 5$. Group HMOD achieves lower entropy than group FSL in both cases. This suggests that the circles generated by the hierarchical modularity-based method are taken more holistically into consideration than Facebook smart lists by the participants to make visibility decisions. In other words, it is more often that a circle in the HMOD group, than that in the FSL group, is marked unanimously as the people who "can see" or "cannot see" a post. We can also see that raising α level indeed increases the average entropy scores in both groups, but the increase is more apparent in group FSL ($\approx 22\%$) than in group HMOD ($\approx 10\%$).

We test the statistical significances of the differences between the entropies from HMOD and FSL. It is however, less straightforward to compare the two, because the entropy scores yielded in group FSL are from a different set of participants, with a different set of EOSN and posts. Nevertheless, it is possible to perform an approximate comparison by a pessimistic pair-wise matching. We first calculate the pair-wise squared entropy differences between FSL and HMOD/MOD, deriving a cost matrix, with which, we match the entropies in the two groups via the linear assignment [43] to minimize the sum of the pair-wise differences (so as to minimize the difference between the two models). Based on the resulting pair-wise matches, we perform the t-tests. It then follows that, in comparing HMOD and FSL, the t-statistic is 9.146 for $\alpha > 1$ and 12.810 for $\alpha > 5$. The t-statistics reject the corresponding null hypotheses with two-tail Confidence Interval (CI) $= 99.9\%$ and one-tail $CI = 99.95\%$. It is then evident that HMOD is significantly better than FSL.

From this user study, we gain more insight into users' privacy decision-making process in online posting. *First*, it is evident that categorical thinking is used when users make binary visibility decisions. *Second*, graph-modularity-based friend communities assist users more efficiently for such decisions than the profile-attribute-based Facebook smart lists. This implies that the former produces the communities that fit the categories of friends that a user has in mind, more than the latter. We examine another state-of-the-art community detection algorithm for EOSN in comparison with the modularity-based algorithm in Section 4 and discuss the implications of the results. Third, the essence of categorical thinking is to reduce the cognitive load. If there are too many information objects (in our case, online friends), hierar-

chical categorization supports the users' visibility-decision-making. When the resolution limit of MOD is reached, the sub-circles can be of more help for the users. The results also give us guidance in designing information visualization systems – categorization and abstraction are important for users to process large amount of information.

4 Community Detection and Social Groups

In this section, we first introduce the two community detection methods of interest (Section 4.1), then describe the datasets (Section 4.2) on which the two algorithms run. We compare and discuss the performances of the two algorithms on these datasets (Section 4.3), and propose an extension of one of the algorithms to accommodate the overlapping nature of online friend groups (Section 4.4). We summarize and compare the performances of all three algorithms by the end of this section.

4.1 Two Models for Community Discovery

From our preliminary user study, we know that, in order to make sense of the friends in one's online social life, it is important to categorize them, either for the ease of processing and memorizing friends' information, or as an efficient means for making decisions. Given the large number of friends that one usually has in EOSN, automated community detection can be very helpful not only as the basis for visibility decisions, as investigated in the previous section, but also for other tasks in online contact management, such as simply keeping an overview, sorting incoming messages, etc. In this section, we examine community detection algorithms and their relationship with real-life social groups. We compare two models for community detection in EOSN: the graph-modularity-based model (MOD) using eigenvalue decomposition [45] and the Generative Model for Friendships (GMF) [39].

Modularity is the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random [46]. Larger modularity value suggests more obvious community structure in the graph. There exists abundant and different techniques that optimize the modularity of a graph. We chose to implement Newman's spectral optimization algorithm that iteratively bisects a given graph using the eigenvectors of the modularity matrix. This approach is generally more accurate than the techniques such as greedy methods and external optimization, and less computationally expensive than global optimization approaches such as simulated annealing [21]. We also implement vertex-moving to improve the final modularity score, as proposed in [46]. Intuitively, in each bisection of the input (sub-)graph, "vertex-moving" moves one vertex at a time, from one (sub-)community to the other, if the modularity is increased, it makes this move permanent. The average, combined complexity of this algorithm is $O(N^2 log\ N)$.

GMF is a recently proposed community detection model that leverages both the friend-profile features and the friend-graph structure in an EOSN [39]. The resulting communities have the following properties: (1) the friends in the same communities have common features, such as education, work; (2) different communities may emphasize different features; (3) the communities may overlap. GMF has been evaluated against the ground-truth communities from three EOSN datasets (as described in Section 4.2), and compared with eight baseline models – Mixed Membership Stochastic Block Models, Block-LDA, K-means clustering, Hierarchical Clustering, Link Clustering, Clique Percolation, Low-Rank Embedding and Multi-Assignment Clustering (as elaborated in [39]). It was demonstrated that GMF generated more accurate communities than the baselines.

4.2 Three EOSN Datasets

The three datasets were collected from Facebook, Twitter and Google+, which are available online[7]. We downloaded these datasets, removed empty files, and discarded the ego-networks whose ground-truth circle(s) contains just one friend. Finally we obtained 10, 909 and 129 ego-networks from Facebook, Twitter and Google+ respectively, which we use for our experiments. Note the data is a subset of the data used in [39]. Each ego-network includes the user's and the friends' profiles, the friend graph and the set of manually constructed circles by the user. For the Twitter and Google+ friend graphs, we ignore their directivity as MOD runs on undirected graphs. We denote the complete set of friends as V, the friend nodes retrieved from the user's ground-truth circles in an EOSN as $V_{circles}$, the friend nodes retrieved from the user's friend graph as V_{edges}, a ground truth circle as c, the set of ground-truth circles as C, an algorithm-generated circle as c' and a set of algorithm-generated circles as C'. The three datasets are summarized in Table 3. We see that $\overline{|V_{circles}|} < \overline{|V_{edges}|}$ for the three datasets, since $V_{edges} \subseteq V$, it indicates that $V_{circles} \subset V$. Moreover, we observe that overlapping ground-truth circles are common, but also limited such that a friend is usually assigned to less than two circles.

4.3 Performances of GMF and MOD

We follow the same method and metrics in [39] to evaluate how well a set of generated circles C' match the user's manual circles C. Balanced Error Rate (BER) [12] and F1 scores are used to measure the matches of circles, as defined in Equation 2 and 3. We use $RBER(c,c')$ to refer to $1 - BER(c,c')$. In order to determine which $c' \in C'$ corresponds to which $c \in C$, we perform a linear assignment using the Hungarian Algorithm [43] to maximize the sum of the pair-wise $RBER$ or $F1$.

[7] https://snap.stanford.edu/data/index.html#socnets [Accessed on Dec 9, 2013]

Table 3 Three ego-network datasets summarized, from left to right: $\overline{|V_{circles}|}$ is the average number of friends from a user's ground-truth circles, $\overline{|V_{edges}|}$ is the average number of friends from a user's friend graph, $\overline{|C|}$ is the average number of a user's ground-truth circles, $\overline{|c|}$ is the average ground-truth circle-size, $\overline{No.Comms.P}$ is the average number of ground-truth circles to which a friend belongs.

EOSN	$\overline{\lvert V_{circles}\rvert}$	$\overline{\lvert V_{edges}\rvert}$	$\overline{\lvert C\rvert}$	$\overline{\lvert c\rvert}$	$\overline{No.Comms.P}$
Facebook (10)	298	423	19.3	26	1.6
Twitter (909)	36	134	4.4	12	1.4
Google+ (129)	304	1948	3.6	135	1.6

$$BER(c,c') = \frac{1}{2}\left(\frac{|c \backslash c'|}{|c|} + \frac{|c' \backslash c|}{|V_{circles}| - |c|}\right) \tag{2}$$

$$F1(c,c') = 2\frac{|c \cap c'|}{|c| + |c'|} \tag{3}$$

We ran GMF[8] and MOD on the ego-networks that only included the friend nodes from ground-truth circles, so that we could compare C and C'. The reason that we ran GMF again instead of directly using its original result was because of the incomplete ego-network data that we could download and some trivial data (e.g. an ego-network containing only one friend) that we discarded afterwards. As such, both GMF and MOD were run on the subsets of the ego-networks that were described in [39], namely 10 Facebook, 909 Twitter and 129 Google+ ego-networks instead of 10, 1000 and 133 ego-networks. Due to the complexity of the algorithm (with the worst case complexity $O(N^3)$, N being the number of friend nodes in an ego-network), we ran GMF for each ego-network with selective K values (the number of communities), $K = 3, 5, 7$ and 9 respectively. Then we select the K value that corresponds to the highest average $\overline{RBER}$ or $\overline{F1}$, and match C and C' for each ego-network via linear assignment. As for MOD, K is automatically derived in the process of modularity maximization. The results are summarized in Table 4 and 5. Note that while certain K of GMF achieves the highest $\overline{RBER}$, it does not necessarily mean this K corresponds to the highest $\overline{F1}$. Thus we have two different sets of combinations of Ks with respect to the $\overline{RBER}$ and $\overline{F1}$ measures. The columns $\overline{|C'|}$ and $\overline{No.Comms.P}$ in Table 5 are based on the average values of these two sets of Ks.

We denote the GMF algorithm that was run on the original ego-network datasets, with a full range of K values checked, as GMF0. This is to differentiate it from the GMF model that we ran on the subsets, with the four K values checked. From Table 4, we notice that the $RBER$ and $F1$ scores of GMF on the Facebook and Google+ datasets are smaller than those of GMF0, and the $RBER$ and $F1$ scores of GMF on the Twitter dataset are comparable to or higher than those of GMF0. The relatively

[8] The code can be downloaded from the author's web page: http://i.stanford.edu/ julian/. We used the default parameters in the code with different K values.

Table 4 The comparison between the results of GMF running on the subsets with four K choices (white columns) and the original sets (gray columns) of the ego-networks: Facebook (Fb), Twitter (Tw) and Google+ (Gp).

GMF	Fb(10)	Fb(10)	Tw(909)	Tw(1000)	Gp(129)	Gp(133)
$\overline{RBER}$	0.83	**0.84**	**0.77**	0.70	0.65	**0.72**
$\overline{F1}$	0.53	**0.59**	0.32	**0.34**	0.24	**0.38**

Table 5 The results of running GMF and MOD on the three subsets of ego-networks. The gray sub-columns are the results for GMF, the white ones are for MOD. $\overline{|C'|}$ is the average number of generated circles, $\overline{|c'|}$ is the average size of each generated circle, $\overline{No.Comms.P}$ is the average number of circles to which each friend belongs.

EOSN	$\overline{RBER}$		$\overline{F1}$		$\overline{\lvert C'\rvert}$		$\overline{\lvert c'\rvert}$		$\overline{No.Comms.P}$	
Facebook	0.83	**0.86**	0.53	**0.67**	3.3	7.0	90	41	1.5	1
Twitter	0.77	**0.81**	0.32	**0.68**	5.2	3.0	7	12	2.7	1
Google+	0.65	**0.75**	0.24	**0.62**	6.9	3.1	44	98	3.8	1

large performance difference on Google+ is due to the limited choices of K in GMF. From Table 5, we see that MOD fully outperforms GMF on $RBER$ and $F1$ measures.

4.4 Multi-membership Modularity-Based Method

From Table 5, we can also see that MOD generates the $\overline{|C'|}$ that is closer to the ground-truth as shown in Table 3. We also know that though overlapping circles are common in the ground-truth, one friend is rarely put into more than two circles, whereas GMF on Twitter and Google+ generates the circles that have $\overline{No.Comms.P}$ equal to or larger than three, which led to its relatively low performance on these datasets. However, a significant limitation of MOD is that it produces non-overlapping communities, while it is obvious that OSN users construct overlapping circles by themselves. Thereby, we propose an extension of MOD that allows multiple circle memberships, which we call Multi-membership Modularity-based community detection, shortly as MMOD. We define a metric we call the External Belongingness (EB as in Equation 4), in which $neighbors(v,c')$ is the number of neighbors (one hop away on the friend graph) of a given friend v in an external circle c', $degree(v)$ is the degree of v. c' is external to v if $v \notin c'$. We first run MOD to derive a set of non-overlapping circles. Then for each friend, we obtain a list of external circles (the circles to which the friend does not belong) with the corresponding EB scores. We subsequently check the highest EB score for each friend, if it exceeds the previously defined θ_{EB}, the friend is assigned to the corresponding external circle. In this way, we obtain a set of overlapping circles with some friends belonging to two circles. However, it remains the question of how to select θ_{EB}. We

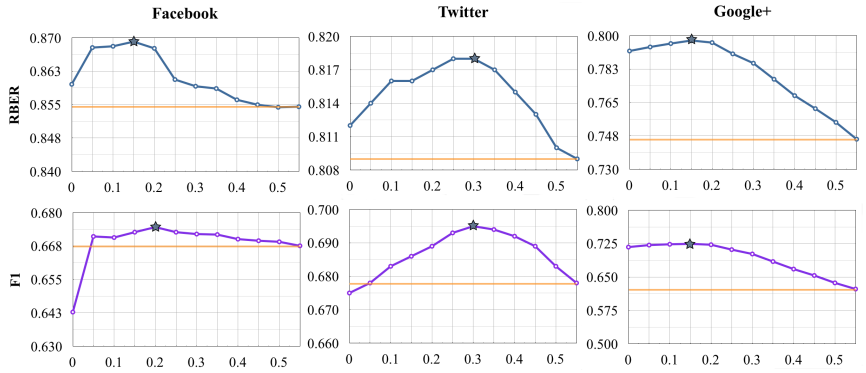


Fig. 6 The *RBER* and $F1$ performances of MMOD with different θ_{EB} values. The baselines are drawn to indicate the corresponding MOD performances and the stars are to mark the optimal θ_{EB} points.

run MMOD with different $\theta_{EB} \in [0, 0.5]$ with the step size 0.05. Then we match the respective overlapped C' with C, the performances are plotted in Figure 6. In each plot of Figure 6, the last point is the average *RBER* or F1 score from MOD, through which a straight horizontal line is drawn to indicate baseline performance. The point with the highest performance is marked with a star.

$$EB(v,c') = \frac{neighbors(v,c')}{degree(v)}, v \in V, v \notin c' \tag{4}$$

From Figure 6, we can see that the performances of MMOD are generally better than those of MOD. The curves also follow the similar trend that increases till some particular θ_{EB} and drops. Around $\theta_{EB} = 0.5$, rarely any friend nodes can be found in external circles, thus the performances regress to be close to MOD's. We also find that the optimal threshold θ_{opt} values for Facebook and Google+ data are similar, which stay around 0.15 for both *RBER* and $F1$, whereas for Twitter data, this value is 0.35. The *RBER* and $F1$ scores of MMOD at these θ_{opt}, along with other results (the same columns as Table 5) are summarized in Table 6, from which we see that MMOD fully outperforms MOD, and that the $\overline{No.Comms.P}$ values are very close to those of the ground-truth datasets. The better results on MMOD also have the implication that people indeed tend to put the friends who are the connectors or hubs in the ego-network into different circles at the same time. We summarize the performances of GMF, MOD and MMOD in Figure 7.

We also observe that θ_{opt} empirically correlates with the average size $\overline{|V_{circles}|}$ of an ego-network, which is around 300 on Facebook and Google+, and 30 on Twitter. For instance, we can describe this relation with Equation 5. If we consider the MMOD-generated circles match the user's manual circles better (indeed, the *RBER* rates are close to or well above 0.8), the relation in Equation 5 suggests that, on the one hand, users tend to manually create less overlapped circles when they have fewer friends. On the other hand, θ_{opt} decreases exponentially slower than the num-

Table 6 The results of running MMOD on the three subsets of ego-networks. $\overline{|C'|}$ is the average number of generated circles, $|c'|$ is the average size of each generated circle, $\overline{No.Comms.P}$ is the average number of circles to which each friend belongs.

EOSN	$\overline{RBER}$	$\overline{F1}$	$\overline{\lvert C'\rvert}$	$\overline{\lvert c'\rvert}$	$\overline{No.Comms.P}$
Facebook	0.87	0.67	7.0	52	1.3
Twitter	0.82	0.70	3.0	18	1.5
Google+	0.80	0.73	3.1	166	1.7

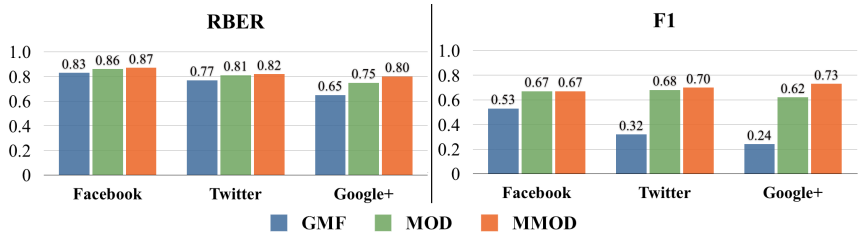


Fig. 7 The overview of the performances of GMF, MOD and MMOD.

ber of one's friends increases, which means that on a relatively large scale (e.g. $|V_{circles}| \in [100, 1000]$), given that EOSN are often sparse [57, 42], users' θ_{opt} for allowing a friend to be in multiple circles remains similar ($\theta_{opt} \in (0.12, 0.20)$ approximately). However, in order to accurately capture the relationship between the number of friends and the optimal threshold, we need a further investigation. It may involve other potentially correlated parameters, more sophisticated models and more data, which is beyond the scope of this work. Equation 5 is manually derived based on the observations from Table 3 and Figure 6. It serves as an intuitive guidance for determining θ_{opt}.

$$|V_{circles}| = 3 \times 10^{(\frac{0.3}{\theta_{opt}})} \Longleftrightarrow \theta_{opt} = \frac{0.3}{lg|V_{circles}| - lg3},\ \theta_{opt} > 0 \tag{5}$$

We perform ANOVA (ANalysis Of VAriance) to compare GMF, MOD and MMOD on the three datasets. The **p** values are summarized in Table 7. We can see that the **p** values on the Facebook dataset are rather high, and the **p** values on the other two datasets are low ($\mathbf{p} < .001$). This means that the variance between the three models is not significant on the Facebook dataset, but very significant on the Twitter and Google+ datasets (in fact, the F-statistics on these two datasets approach the ends of the corresponding F-distribution curves.) In Figure 7, the observed differences were statistically significant for both $\overline{RBER}$ and $\overline{F1}$ on the Twitter and Google+ datasets (all $\mathbf{p} < .001$ for one-way ANOVAs), but not for the Facebook dataset ($\mathbf{p} = .43$ for $\overline{RBER}$ and $\mathbf{p} = .13$ for $\overline{F1}$). The latter may be a result of the small sample.

Table 7 The **p** values of the ANOVA for GMF, MOD and MMOD, of both $\overline{RBER}$ and $\overline{F1}$measures, on the datasets of Facebook, Twitter and Google+ respectively.

	Facebook	Twitter	Google+
$\overline{RBER}$	.43	$< .001$	$< .001$
$\overline{F1}$	.13	$< .001$	$< .001$

4.5 Discrepancy between Predicted and Manual Circles

Though a community discovery algorithm can predict reasonably good circles, it is unlikely that it can make a perfect prediction. This attributes to the fact that manual circle-creation process is inherently subjective, and varies on the same person for different purposes. The ground-truth circles of the ten Facebook users that we used in our experiments were obtained by a Facebook app[9], in which the user entered comma-separated category labels for each friend. Existing labels could be reused by a selection from a drop-down box. Each label represented a circle to which a friend belonged. The text cue for entering the label(s) for each friend **Fr** was "I know **Fr** because ..." followed by the label-entering text-field. In another exercise [15] of friend-grouping, the groups (i.e. circles) were constructed by "card sorting". The name of each friend of a participant's was printed on a paper card. Several cards were randomly selected and spread on a table, the participant was then asked to assign the rest of the cards to the selected ones to form groups. We can see that the Facebook app friend-grouping exercise encourages more overlapping circles to be created than the card-sorting exercise.

Different user interfaces may directly reflect intrinsic and systematic differences on a functional level, rather than on a perceptual level. Facebook provides a social platform mainly for mutual friends – two people become friends when one "accepts" the other's "friend request". The friends of friends are recommended if the user wants to add more friends. Twitter and Google+ implement a "follower-followee" mechanism, which means a friendship is not necessarily reciprocal. On Twitter, the user clicks the "follow" button to follow a "friend", every newly followed friend is not necessarily put into a friend list (i.e. a circle), whereas on Google+, the "follow" button becomes the "add" button, and every newly followed friend has to be added into one of the existing circles or a new one. This is an important reason that the number of friends in Google+ circles is much more than that in Twitter lists, as shown in Table 3. We see that people create circles differently under different circumstances, consciously or unconsciously. It is therefore important to create interfaces that help users gain insights about their EOSN friendships from different aspects, and let them form their own friend circles with more informed decisions.

Moreover, social and cognitive theories shed light on human social grouping behavior and inform computer scientists to design community detection algorithms

[9] https://www.facebook.com/apps/application.php?id=201704403232744 [Accessed on Dec 12, 2013]

and interactive visualizations. The social brain hypothesis (SBH) offers a framework for integrating evolutionary and social psychological perspectives on human social complexity. SBH predicts a natural community size of around 150 for modern humans (Dunbar's number [18]), and now there is considerable evidence confirming that this is the typical size of both personal social networks and key types of human community [17]. Note that 150 is the typical size of a person's active network, in which she knows how these the friends fit into her social world and they know how she fits into theirs [17]. From the literature in cognitive science, we also know that there is the cognitive capacity limit in human Short-Term-Memory (STM), which is inline with the theory of categorical thinking (Section 3.1). This capacity limit is averaging on seven [40], which means that people can remember seven chunks of information in STM tasks. In our case, we can consider a chunk to be a group of friends. This limit is subject to debate, later evidences showed that it was a high estimate, lower numbers were proposed, e.g. four [14]. The theories on social group size and human's cognitive capacity limit provide more incentives for interactive visualizations, which should enable users to flexibly interact with friend visual objects on different granularity-levels – from (sub-)groups of friends to individual friends.

5 Improving the Tool Design

From the previous sections, we understand that grouping friends is important for OSN users to manage online contacts and make privacy decisions. A carefully designed community detection algorithm can produce decent friend circles that match users' manual circles, but this matching is hardly perfect due to the subjective nature of friend grouping. To close the gap between computer-based grouping and human grouping, tools need to be designed and built. The goal of such tool that leverages interactive visualization and accurate community detection is not only to show its users their structured friendships, but more importantly, to make the structures more usable for the users.

We have introduced a tool in our user study to assist users' visibility decision-making (Section 3). It visualizes the generated circles of the user's friends, and allows hierarchical exploration. However, this tool addressed only part of the information about the user's friends, with limited navigation functions. As various taxonomies for visual analytics or information visualization unanimously emphasized [25, 32, 63, 53], presenting multiple aspects and providing multiple perspectives are essential for visualizing large and complex data. We developed a new online application named FreeBu[10]. We motivate and describe three more views that supplement the CircleTree view (Section 3.4). All the four views serve a two-fold purpose: (1) to provide users with different insights about their own ego-networks, (2) users can manually construct their Facebook friend lists with the tool.

[10] http://people.cs.kuleuven.be/ bo.gao/freebu/ [Accessed on Dec 12, 2013]

From Section 3 we have known that FSL are less efficient in visibility decision-making than the communities generated by MOD, suggesting that graph-based communities coincide more with the friend groups that a user has in mind. An investigation (Section 4) in the three community detection algorithms GMF, MOD and MMOD has shown that the ground-truth circles are still in favor of the graph and modularity-based methods. And introducing overlaps further increases the *RBER* and $F1$ accuracies. However, it is incorrect to assume the human friend-grouping process is systematically similar to the algorithmic process just because both produce similar groups. The CircleTree visualization shows circles as friend groups, in which each member is labeled by the name. Other types of data, such as profile, posts, chat history, friend graph, etc. are also potentially useful for the user's understanding of her own ego-network. They can inspire the user to reflect on her online contacts and facilitate friend-group creation.

As the recent study [15] on OSN-friend-grouping shows, people do consider attributes, such as school, music band or youth community, when they group friends, we refer to this type of grouping strategy as the *Attribute* strategy. Also, people indeed tend to put the friends who are mutually friends into the same the group, we refer to this strategy as the *Graph* strategy. Another graph-related, but slightly different grouping strategy is based on some particular friends – "I know those friends via this friend", we refer to it as the *Connection* strategy. The fourth strategy is based on trust or closeness, to which we refer as the *Closeness* strategy. Informed by these grouping strategies, we have four visualizations in FreeBu to accommodate users' comprehension of their online friends and help users create friend groups semi-automatically. The four visualizations/views are described as follows:

- **Circle View** The circle view (i.e. the CircleTree Visualization) is for the *Graph* strategy. Mutually connected friends tend to be put in the same circle, the user can drag and drop a circle or an individual node to compose her own Facebook friend list (Figure 8a). The group circles with different sizes also provide the user with a sense of ordering, helping her quickly find outliers or surprising circles. The visualization and interaction strategies are detailed in Section 3.4.
- **Map View** The map view is for the *Graph* and the *Connection* strategies. The user's friend graph is directly visualized in a force-directed layout with the Fruchterman-Reingold algorithm [23], which is a typical graph-layout algorithm. It pulls connected nodes together and pushes disconnected nodes apart. Users can easily observe visual clusters and hub-nodes. The user can zoom and pan to explore the graph, zooming-in brings out the node labels. Mouse-hover on a friend node also brings out the friend's name label, meanwhile highlights the connections of this friend on the graph. Right-clicking a friend node will automatically select this node as well as its neighbors. User can then drag and drop the selected nodes to compose her own friend list (Figure 8c). Furthermore, the nodes' radii are set proportionally to the corresponding Betweenness [44] scores, so that the important nodes that connect different parts of the user's ego-network are enlarged and emphasized. It has been shown that the bridging structure in a user's EOSN is important for predicting strong social ties, such as romantic partners [4]. To make the group-creation more flexible, the point-in-polygon function is

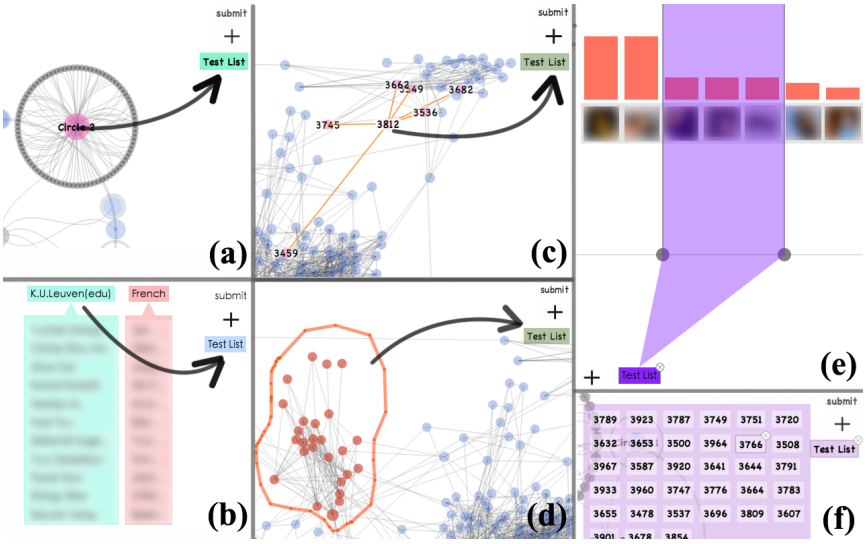


Fig. 8 This figure shows the four views in FreeBu and the drag-drop actions to compose user-defined lists. Each arrow indicates a group-level drag-drop action.

implemented. The user can turn on this function by pressing the "pen" button on the bottom-right corner of the canvas and draw a polygon to enclose and select the nodes of interest, and drag-drop the selected nodes to compose lists (Figure 8d).

- **Column View** The column view is for the *Attribute* strategy. The column view generates the groups of friends based on common profile-attributes between friends, which is a generalization of Facebook smart lists. Each column represents a group. The "head" of the column is labeled with the corresponding attribute-value name. The "body" is a stack of friend name tags belonging to that column. If a column contains more than N_{col} (e.g. $N_{col} = 12$), only N_{col} friend tags are initially shown in the body of the column, with the "..." symbol to indicate there is more tags. Mouse-hover on the head of a column expands the column and show all the member names. The heights of the columns are proportional to corresponding the numbers of friends. Users can scroll left or right with mouse wheel to explore the columns. They can click the "overview" button for a summary of all the column labels. The user can drag and drop a column or an individual tag to compose lists (Figure 8b). Moreover, the user can drag and drop columns into the "intersection" area at the bottom of the canvas. This area keeps the members that satisfy the attribute values from the columns. The user can then use intersected area (also via drag-drop) to compose her friend lists.
- **Rank View** The rank view is for the *Closeness* strategy. Studies [56, 17] have shown that interaction frequency linearly corresponds to the strength of interpersonal ties. We visualize the users' friends by aligning their profile photos horizontally near the middle of the canvas. The photos are ranked according to the

communication frequencies of the user with her friends in Facebook chat. On top of each photo, a bar is shown if there is a communication history of the user with that friend. The more frequently the user chatted with a friend, the higher the bar is. The user can scroll left or right with mouse wheel to see the bars and photos. Mouse-hover can enlarge a photo can brings out the name beneath it. The user can select one or more friends by moving the two "knobs" with vertical lines. Clicking on a user-defined list "absorbs" the friends that are "clipped" by the two knobs into that list (Figure 8e).

The four views share a similar way for creating customized friend lists. The user starts by clicking the "plus" button to add a new, empty list, aligned on the right (in the first three views) or the bottom (in the rank view) of the canvas. Each list is shown as a rectangle. The user can right-click a list to edit its name. Drag-drop actions put selected friends into a list, as illustrated in Figure 8a-d, whereas in the rank view, "clipped" friends are put in a list by user clicking on the list, as shown in Figure 8e. Mouse-hover on a list brings out the friend-name tags of the list in a grid layout. Mouse-hover on a list or a tag also brings out the "remove" button, as shown in Figure 8f. In this way, the user can remove a list or a member if needed. The user can submit the lists to her Facebook account by clicking the "submit" button.

An elaborate multi-method user study on the usefulness and perceived values of FreeBu is beyond the scope of this chapter. We refer to [16] that has detailed such study. Through a factor analysis, it showed that FreeBu received high scores (between 4 and 6 on a 7-point Likert Scale) on several factors of perceived values. These factors include Audience Control and Audience Reflection. The first factor refers to sharing information with differentiated friends. For example, "FreeBu helps me create Facebook friend lists". The second factor refers to the reflection and re-evaluation of one's friends in her EOSN. For example, "FreeBu clarifies my relationships with others of whom I am not fully aware". The regression analyses in [16] further identified several attributes that directed users' attention and guided users' usage of the tool. For example, in the map view of FreeBu, users are more interested in the friends who act like hubs (with high betweenness scores) or the friends who are outliers (with low degree scores) in their social networks. In the rank view, users were very interested in the friends to whom they often communicated.

6 Conclusion

In this section, we first summarize our research, then address the future work.

6.1 Summary

In this work, we addressed the issue related to privacy-decision-making in Online Social Networks (OSN). The available large amount of information about friends

overwhelms a user. It is then difficult for the user to decide the audience for her online posts. Various research work has pointed to friend categorization. Indeed, the theories in categorical thinking and social networking limits provide us with further support. Leveraging humans' innate ability to process visual information, we developed an online visualization application to help users explore and group friends. It requires careful design choices in both visualizations and algorithms. We first reviewed various existing tools, identified their merits and limits. We then described our first tool based on the CircleTree visualization and the modularity-based community detection (MOD). The former is our new visualization design. We conducted a user study to investigate OSN users' visibility-decision performances with two different grouping methods, under the CircleTree visualization. The participants were divided into two groups, one used hierarchical, modularity-based community detection method (HMOD) interactively, the other used Facebook smart lists (FSL). We found that the former group of participants utilized the circles more efficiently the latter. This provides the evidence that HMOD is more supportive than FSL for visibility decisions. It also suggests that graph-based algorithms can produce the communities that match users' manual circles, more than attribute-based ones.

We then compared MOD with another community detection model, Generative Model for Friendships (GMF). It had been shown that GMF outperformed the other eight community detection models [39]. The corresponding nine algorithms were run on three ego-network datasets, and compared to ground-truth circles. We ran MOD and GMF on the sub-datasets (due to the availability of the data), and found that MOD outperformed GMF. We also examined the characteristics of the ground-truth circles and proposed the Multi-membership Modularity-based community detection method (MMOD) that produced overlapping communities, with similar overlapping rate to the ground-truth. We then found that MMOD outperformed MOD.

It is important to note that improving community detection algorithms alone is insufficient. Users need informative visualizations to comprehend her online friends and construct her own friend lists. Guided by relevant sociological research and visualization design taxonomies, we developed three more interactive visualizations that compensated the CircleTree visualization. The four visualizations are based on four different friend-grouping strategies. They incorporate similar list-construction user interfaces.

In summary, this work begins with the concerns for online privacy and contact management, results a web application for EOSN friend-exploration/grouping. We examined in detail the design choices from different perspectives: information visualization, community detection algorithms, human cognition for visual perception and information processing, and social theory on social groups.

6.2 Limitations and Outlook

We identify several main improvements for FreeBu:

- In the four views, each friend is only represented by her name (the rank view also includes photos). More information, such as photo, profile, recent status and likes can be summarized in an "info box" that appears besides each focused friend.
- There often exist the friends who do not connect to other friends in an ego-network. The loners can be randomly mixed into the circles in the circle view or scattered in the force-directed graph layout in the map view. It is then more orderly to collect these loners into the same circle or to map them in proximity in the graph.
- We can improve the circle view by applying MMOD (Section 4.4).
- For the circle view, we notice that the user needs to zoom-in fairly deeply to reveal the friend names in a circle. This can be improved by modifying the label-revelation threshold. Also, the positioning of the name labels needs adjustment, so as to avoid overlaps, while maintaining a grouping structure.
- The graph layout in the map view can be colored according to the communities detected by MOD, similar to InMaps (Section 2). The drawback of discretizing community colors is that it ignores the continuity of the friend graph. Some friends are meant to be community-ambiguous. One way to address this issue is via gradient colors. First, a friend's membership to the circle is characterized some measure, e.g. its clustering coefficient (as that in Social Graph in Section 2). However, we need to be more careful to make people perceive such fusion as a natural transition between communities on the graph.
- Because zooming can create too much local focus and lose global context, it could be more helpful for users to add the "fisheye-view" [24] and "map-window" [7] functions in the circle and map views.
- In all the four views, we can add filtering and searching function to improve users' exploratory experience.
- FreeBu users have reported in some cases rendering visualizations is slow. Complex visualizations and user interactivity occupy a large part of browser resources, sometimes result slow response or crash. Though the current standard web technologies are encouragingly evolving, such as improved graphics rendering capabilities in HTML5, faster built-in Javascript engines, the browser-based computation power is still limited for large-scale, online, interactive visualizations. For tools like FreeBu, visualization programs need to be more economic.

Acknowledgement

We thank the Flemish Agency for Innovation through Science and Technology (IWT) and the Fonds Wetenschappelijk Onderzoek – Vlaanderen (FWO) for support through the projects SPION (grant number 100048) resp. Data Mining for Privacy in Social Networks (grant number G068611N).

References

[1] Acquisti A, Gross R (2006) Imagined communities: Awareness, information sharing, and privacy on the facebook. In: Privacy enhancing technologies, Springer, pp 36–58

[2] Allport GW (1979) The Nature of Human Prejudice. Basic books

[3] Auber D, Mary P (2013) Tulip – an information visualization framework dedicated to the analysis and visualization of relational data. URL (Weblog): http://tuliplabrifr/TulipDrupal/ (Download: 1129 2013)

[4] Backstrom L, Kleinberg J (2014) Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on facebook. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, ACM, pp 831–841

[5] Bakewell S (2010) How to live: A life of Montaigne in one question and twenty attempts at an answer. Random House

[6] Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. In: ICWSM

[7] Beard D, Walker J (1990) Navigational techniques to improve the display of large two-dimensional spaces. Behaviour & Information Technology 9(6):451–466

[8] Bernstein MS, Bakshy E, Burke M, Karrer B (2013) Quantifying the invisible audience in social networks. In: ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2013).

[9] Bisson G, Blanch R (2012) Improving visualization of large hierarchical clustering. In: Information Visualisation (IV), 2012 16th International Conference on, IEEE, pp 220–228

[10] Boyd DM (2008) Taken out of context: American teen sociality in networked publics. ProQuest

[11] Bruls M, Huizing K, Van Wijk JJ (2000) Squarified treemaps. In: Data Visualization 2000, Springer, pp 33–42

[12] Chen YW, Lin CJ (2006) Combining svms with various feature selection strategies. In: Feature Extraction, Springer, pp 315–324

[13] Collins AM, Quillian MR (1969) Retrieval time from semantic memory. Journal of verbal learning and verbal behavior 8(2):240–247

[14] Cowan N (2001) The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and brain sciences 24(1):87–114

[15] De Wolf R, Pierson J (2013) Whos my audience again? understanding audience management strategies for designing privacy management technologies. Telematics and Informatics

[16] De Wolf R, Gao B, Berendt B, Pierson J (2015) Interactive grouping technology for social network sites: exploring users' perceived values and actual behaviours, submitted for publication, at the ACM conference on Computer-Supported Cooperative Work and Social Computing

[17] Dunbar R, Sutcliffe A (2012) Social complexity and intelligence. The Oxford Handbook of Comparative Evolutionary Psychology p 102

[18] Dunbar RI (1992) Neocortex size as a constraint on group size in primates. Journal of Human Evolution 22(6):469–493
[19] Fang L, Kim H, LeFevre K, Tami A (2010) A privacy recommendation wizard for users of social networking sites. In: Proceedings of the 17th ACM conference on Computer and communications security, ACM, pp 630–632
[20] Field DJ, Hayes A, Hess RF (1993) Contour integration by the human visual system: Evidence for a local association field. Vision research 33(2):173–193
[21] Fortunato S (2010) Community detection in graphs. Physics Reports 486(3):75–174
[22] Fortunato S, Barthelemy M (2007) Resolution limit in community detection. Proceedings of the National Academy of Sciences 104(1):36–41
[23] Fruchterman TM, Reingold EM (1991) Graph drawing by force-directed placement. Software: Practice and experience 21(11):1129–1164
[24] Furnas GW (1981) The fisheye view: A new look at structured files. Tech. rep., Bell Laboratories Technical Memorandum
[25] Furnas GW (1986) Generalized fisheye views, vol 17. ACM
[26] Gao B, Berendt B, Clarke D, Wolf RD, Peetz T, Pierson J, Sayaf R (2012) Interactive grouping of friends in osn: Towards online context management. In: ICDM Workshops, IEEE Computer Society, pp 555–562
[27] Ghoniem M, Fekete JD, Castagliola P (2004) A comparison of the readability of graphs using node-link and matrix-based representations. In: Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on, IEEE, pp 17–24
[28] Green-Armytage P (2010) A colour alphabet and the limits of colour coding. JAIC-Journal of the International Colour Association 5
[29] Grivet S, Auber D, Domenger JP, Melancon G (2006) Bubble tree drawing algorithm. In: Computer Vision and Graphics, Springer, pp 633–641
[30] Gürses S (2010) Multilateral privacy requirements analysis in online social network services. PhD thesis, PhD thesis, Department of Computer Science, KU Leuven
[31] Healey CG, Enns JT (1999) Large datasets at a glance: Combining textures and colors in scientific visualization. Visualization and Computer Graphics, IEEE Transactions on 5(2):145–167
[32] Heer J, Shneiderman B (2012) Interactive dynamics for visual analysis. Queue 10(2):30
[33] Johnson B, Shneiderman B (1991) Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In: Visualization, 1991. Visualization'91, Proceedings., IEEE Conference on, IEEE, pp 284–291
[34] Kilmer R, Kilmer WO (2014) Designing interiors. John Wiley & Sons
[35] Lee C, Cunningham P (2013) Community detection: effective evaluation on large social networks. Journal of Complex Networks p cnt012
[36] Macrae CN, Bodenhausen GV (2000) Social cognition: Thinking categorically about others. Annual review of psychology 51(1):93–120
[37] Marwick AE, boyd dm (2011) I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media & Society 13(1):114–133

[38] Mazzia A, LeFevre K, Adar E (2012) The pviz comprehension tool for social network privacy settings. In: Proceedings of the Eighth Symposium on Usable Privacy and Security, ACM, p 13
[39] McAuley JJ, Leskovec J (2012) Learning to discover social circles in ego networks. In: NIPS, pp 548–556
[40] Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review 63(2):81
[41] Moore K, McElroy JC (2012) The influence of personality on facebook usage, wall postings, and regret. Computers in Human Behavior 28(1):267–274
[42] Moreira AA, Paula DR, Costa Filho RN, Andrade Jr JS (2006) Competitive cluster growth in complex networks. Physical Review E 73(6):065,101
[43] Munkres J (1957) Algorithms for the assignment and transportation problems. Journal of the Society for Industrial & Applied Mathematics 5(1):32–38
[44] Newman ME (2005) A measure of betweenness centrality based on random walks. Social networks 27(1):39–54
[45] Newman ME (2006) Finding community structure in networks using the eigenvectors of matrices. Physical review E 74(3):036,104
[46] Newman ME (2006) Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23):8577–8582
[47] Nguyen QV, Huang ML (2002) A space-optimized tree visualization. In: Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on, IEEE, pp 85–92
[48] Raynes-Goldie K (2010) Aliases, creeping, and wall cleaning: Understanding privacy in the age of facebook. First Monday 15(1)
[49] Rodrigues EM, Milic-Frayling N, Smith M, Shneiderman B, Hansen D (2011) Group-in-a-box layout for multi-faceted analysis of communities. In: Privacy, security, risk and trust (passat), 2011 IEEE third international conference on and 2011 ieee third international conference on social computing (socialcom), IEEE, pp 354–361
[50] Seo J, Shneiderman B (2002) Interactively exploring hierarchical clustering results [gene identification]. Computer 35(7):80–86
[51] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13(11):2498–2504
[52] Shiffrin RM, Schneider W (1977) Controlled and automatic human information processing: Ii. perceptual learning, automatic attending and a general theory. Psychological review 84(2):127
[53] Shneiderman B, Dunne C (2013) Interactive network exploration to derive insights: filtering, clustering, grouping, and simplification. In: Graph Drawing, Springer, pp 2–18
[54] Shneiderman B, Wattenberg M (2001) Ordered treemap layouts. In: Proceedings of the IEEE Symposium on Information Visualization 2001, vol 73078
[55] Stevens SS (1957) On the psychophysical law. Psychological review 64(3):153

[56] Sutcliffe A, Dunbar R, Binder J, Arrow H (2012) Relationships and the social brain: Integrating psychological and evolutionary perspectives. British journal of psychology 103(2):149–168

[57] Ugander J, Karrer B, Backstrom L, Marlow C (2011) The anatomy of the facebook social graph. arXiv preprint arXiv:11114503

[58] Van Wijk JJ, Van de Wetering H (1999) Cushion treemaps: Visualization of hierarchical information. In: Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on, IEEE, pp 73–78

[59] Wang L, Giesen J, McDonnell KT, Zolliker P, Mueller K (2008) Color design for illustrative visualization. Visualization and Computer Graphics, IEEE Transactions on 14(6):1739–1754

[60] Wang Y, Norcie G, Komanduri S, Acquisti A, Leon PG, Cranor LF (2011) I regretted the minute i pressed share: A qualitative study of regrets on facebook. In: Proceedings of the Seventh Symposium on Usable Privacy and Security, ACM, p 10

[61] Ware C, Purchase H, Colpoys L, McGill M (2002) Cognitive measurements of graph aesthetics. Information Visualization 1(2):103–110

[62] Wolfram S (2013) Data science of the facebook world. URL http://blog.stephenwolfram.com/2013/04/data-science-of-the-facebook-world/, retrieved Nov 30, 2013

[63] Yi JS, ah Kang Y, Stasko JT, Jacko JA (2007) Toward a deeper understanding of the role of interaction in information visualization. Visualization and Computer Graphics, IEEE Transactions on 13(6):1224–1231

[64] Zuckerberg M (2012) One Billion People on Facebook. URL http://newsroom.fb.com/news/2012/10/one-billion-people-on-facebook/, retrieved Jul 19, 2014

5

PAPER 3:

Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence

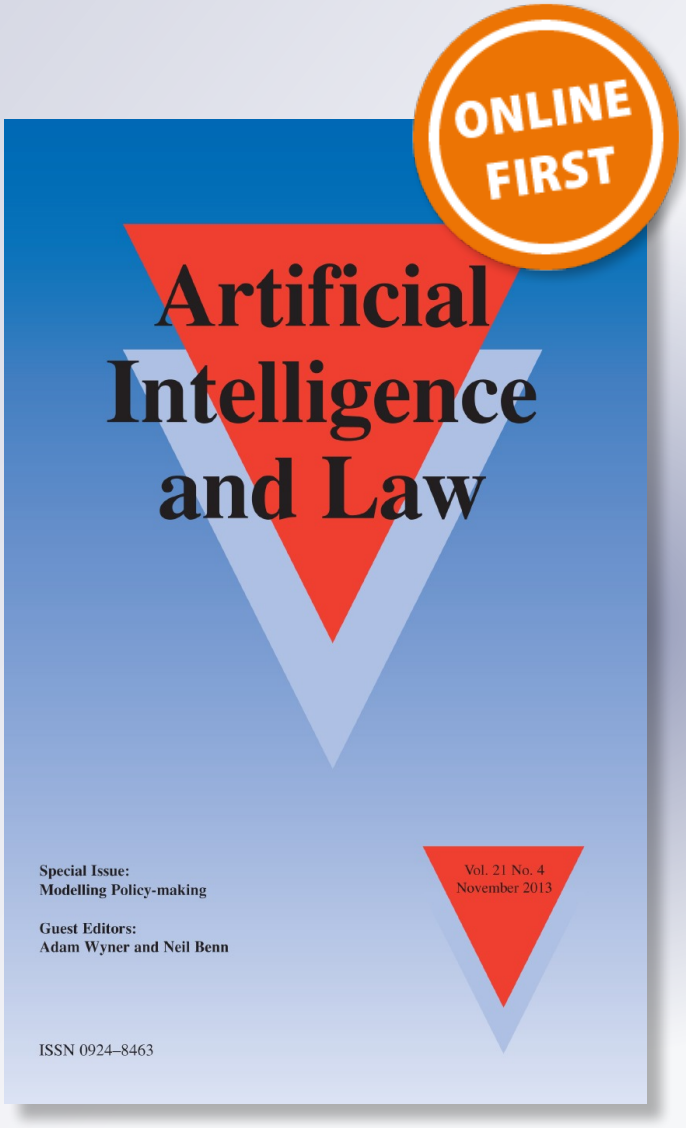
Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence

Bettina Berendt & Sören Preibusch

Artificial Intelligence and Law

ISSN 0924-8463

Artif Intell Law
DOI 10.1007/s10506-013-9152-0



Springer

Artif Intell Law
DOI 10.1007/s10506-013-9152-0

Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence

Bettina Berendt · Sören Preibusch

Abstract Decision makers in banking, insurance or employment mitigate many of their risks by telling "good" individuals and "bad" individuals apart. Laws codify societal understandings of which factors are legitimate grounds for differential treatment (and when and in which contexts)—or are considered unfair discrimination, including gender, ethnicity or age. Discrimination-aware data mining (DADM) implements the hope that information technology supporting the decision process can also keep it free from unjust grounds. However, constraining data mining to exclude a fixed enumeration of potentially discriminatory features is insufficient. We argue for complementing it with exploratory DADM, where discriminatory patterns are discovered and flagged rather than suppressed. This article discusses the relative merits of constraint-oriented and exploratory DADM from a conceptual viewpoint. In addition, we consider the case of loan applications to empirically assess the fitness of both discrimination-aware data mining approaches for two of their typical usage scenarios: prevention and detection. Using Mechanical Turk, 215 US-based participants were randomly placed in the roles of a bank clerk (discrimination prevention) or a citizen / policy advisor (detection). They were tasked to recommend or predict the approval or denial of a loan, across three experimental conditions: discrimination-unaware data mining, exploratory, and constraint-oriented DADM (eDADM resp. cDADM). The discrimination-aware tool support in the eDADM and cDADM treatments led to significantly higher proportions of correct decisions, which were also motivated more accurately. There is significant evidence that the relative advantage of discrimination-aware techniques depends on their intended usage. For users focussed on making and motivating their

B. Berendt (✉)
Department of Computer Science, KU Leuven, Leuven, Belgium
e-mail: bettina.berendt@cs.kuleuven.be

S. Preibusch
Microsoft Research, Cambridge, UK
e-mail: spr@microsoft.com

Published online: 10 January 2014

decisions in non-discriminatory ways, cDADM resulted in more accurate and less discriminatory results than eDADM. For users focussed on monitoring for preventing discriminatory decisions and motivating these conclusions, eDADM yielded more accurate results than cDADM.

Keywords Discrimination discovery and prevention · Data mining for decision support · Discrimination-aware data mining · Responsible data mining · Evaluation · User studies · Online experiment · Mechanical Turk

1 Introduction

In our computer-mediated lives, data supports decisions and carries value that promises unprecedented levels of convenience. The insights that can be inferred from large datasets are however not immediately accessible. They require processes of "knowledge discovery" (Shearer 2000). Knowledge discovery comprises the statistical analysis of data with the help of data mining methods. It also encompasses pre-processing and deployment, as well as the human expertise driving these sub-processes, as integral parts. Many Web users have already profited from data mining in recommender systems, which support their consumption choices or search queries. But data mining is also used when designing HIV vaccines (Heckerman 2013) or with the aim of keeping cities safe (Microsoft 2012). In e-Commerce, banking, insurance, or employment, data mining is often used to segregate "good" from "bad" individuals (Boston Consulting 2012; Duhigg 2009). Besides promising economic advantages, this raises questions of discrimination, not only within the organisations deploying data mining tools, but also among supervisory authorities and social activists.

Differentiation—making a distinction based on some features or attributes—is a fundamental characteristic of human cognition and behaviour. People apply differential treatment to other people, allowing some but not all to vote, applying certain laws to them, giving them jobs, and granting them loans—or denying them the privileges associated with these rights and decisions. Part of the social contract of any society is that certain attributes are accepted for differentiation, while others are not. Non-accepted attributes are those that violate the legal principle of equality, which has found its expression in fundamental and wide-reaching legal codifications such as Article 7 of the Universal Declaration of Human Rights. This article states that "All are equal before the law and are entitled without any discrimination to equal protection of the law." The term 'discrimination' denotes a differentiation on non-accepted grounds. To avoid it, one must treat equal things equally and unequal things unequally. In many countries, individuals are protected by a range of laws against discrimination by the state and also by private actors such as employers. Along with societal notions of what constitutes wanted and unwanted differentiations, the legal demarcations between accepted and non-accepted grounds develop over time, as do the legal groupings of what is equal and what is not.

Many instruments have been proposed for fighting discrimination once it has been outlawed, but discrimination proves to be tenacious. Currently, much hope rests with information technology on which decisions increasingly rely. An appropriately modified algorithm should help to avoid discrimination. In the insurance industry, for instance, data analysis may generate gender-blind tariffs to comply with the new European Union's requirement of unisex policies.

The general *research question* we address in the present article is how to best support the monitoring, understanding, and avoidance of discrimination with the help of information technology. Specifically, we investigate how data mining can act as an instrument against discrimination. We investigate when it is better to hide discriminatory features, and when it is better to reveal and draw attention to them. We also derive recommendations for algorithm and interface design, and discuss the potentials and limitations with regard to further goals such as transparency.

Whether deliberately or unwittingly, discrimination originates in human decisions, which may be tool-supported. Our investigation therefore targets the interface between technology and its human users. We use an *empirical methodology* to quantitatively assess the ability of data mining and the tools displaying its results, to prevent discrimination in decision making. Indeed, deployment and result communication are integral parts of a data mining and knowledge discovery system. We conducted a user study where participants were equipped with data mining solutions to help them make or monitor decisions which could be discriminatory.

Our *contribution* is twofold. First, we critically discuss the emerging area of discrimination-aware data mining (DADM). We argue why the standard approach to DADM is useful and necessary, but also why it falls short of the full technical potential of data mining and also performs sub-par in fighting discrimination. We propose and evaluate a complementary form of DADM, which we call exploratory. Exploratory DADM focusses on revealing and drawing attention to discrimination in data, as opposed to traditional DADM that aims at "hiding" it. We argue that an exploratory approach is needed to find new and unexpected features and patterns of discrimination and is therefore a required complement for effectively avoiding discrimination. As our second contribution, we present empirical evidence to answer the research questions. Using a large-scale experimental user study, we uncover the relative advantages of both forms of DADM in the settings of a bank and an anti-discrimination agency. These correspond to the archetypical applications of data mining in decision support: making and monitoring decisions. To the best of our knowledge, this study represents the first user-centric evaluation of DADM described in the scientific literature; it extends on our previous small-scale exploratory study, which we briefly summarise in this paper.

The remainder of this article is structured as follows: In Sect. 2, we give an overview of related work. In particular, we propose the new classification of DADM approaches and give a brief survey of the literature structured by this framework. In Sect. 3, we discuss appropriate use cases and derive recommendations for DADM evaluation foci. We summarise an exploratory user study ($n = 20$) in which we demonstrated the effectiveness of exploratory DADM in detecting actionable patterns of differentiation and discrimination. Section 4 reports on a new, large-

scale multi-treatment user study ($n = 215$) in which we focussed on the relative advantages of the two forms of DADM in different settings. We conclude with an outlook on future work in Sect. 5.

2 Constraint-orientation versus exploration: a new framework for related work in DADM

To understand the range of DADM, we need to take a step back and ask about the fundamental relations between data mining (discrimination-aware or not) and discrimination (Sect. 2.1). From this, we derive our notion of *constraint-oriented DADM* as a description of most of the current work in the field (Sect. 2.2). While this is a very important approach, it needs to be complemented by *exploratory DADM* (Sect. 2.3).[1]

2.1 Data mining and discrimination

We understand *data mining* in the more general sense of "knowledge discovery" (Shearer 2000) and therefore consider pre-processing and deployment as integral parts. Data mining includes descriptive aspects (when it is used as exploratory data analysis) as well as prescriptive aspects (when it is used for decision support, in recommender systems, etc.).

In a wide sense, *discrimination* is to "make a distinction […] on grounds of [some feature]"; in a narrow sense one "make[s] a distinction, esp. unjustly on grounds of race or colour or sex" (Sykes 1982). Such "unjust" grounds are legally codified in many countries and may include further characteristics. In the following, we will call them *discrimination-indexed attributes/features*.[2] A comprehensive multi-disciplinary overview of discrimination research is provided in Romei and Ruggieri (2014).

Discrimination in the narrow sense may be understood as occurring if and only if one differentiates by such grounds. While *discrimination in the legal sense* often consists of a differentiation in this sense, this is not always the case. It is impossible, within the scope of this article, to describe this notion (in fact, class of notions) exhaustively. Instead, we will highlight important divergences between discrimination in the narrow sense and discrimination in the legal sense, using as an example European (EU) law on discrimination by gender. Where applicable, we will focus on the European "Gender Directive" 2004/113/EC (EU 2004) because its

[1] Sections 2 and 3.1–3.3 extend on a previous workshop paper (Berendt and Preibusch 2012), and Sect. 3.4 summarises the user study presented in detail in that paper.

[2] Otherwise called, e.g., "potentially discriminatory (PD) items" (Pedreschi et al. 2008) or "sensitive attributes" (Hajian and Domingo-Ferrer 2013; Kamiran et al. 2010). A *feature* or *item* is an *attribute* with a value or value range; thus for example "gender" is an attribute and "female" a feature. All three terms refer to the formal representation of *legal grounds* of discrimination (the reasons specified by the law that will serve as a basis for demanding relief) and other grounds in the databases used for data mining. While Pedreschi et al. (2008) point out that PD items may comprise more than just legally-defined sensitive attributes, they still assume a priori knowledge about these items.

application area is closest to the example setting chosen in the experiment described in Sect. 4 below.

- Whether a given differentiation in treatment amounts to discrimination may depend on the agent performing it. States are mainly bound by Article 14 of the European Convention on Human Rights and Articles 18 and 19 of the Treaty on the Functioning of the European Union, private parties in their role as suppliers of goods and services by the national implementations of the "Gender Directive" 2004/113/EC (EU 2004), and private parties in their role as employers by the national implementations of the Equal Treatment Directive 2006/54/EC (EU 2006).
- A differentiation in treatment may amount to discrimination when it is based directly on the discrimination-indexed feature (so-called "direct discrimination"), but discrimination can also result from decisions based on other, seemingly neutral features highly correlated with the discrimination-indexed features (so-called "indirect discrimination"), e.g. EU (2004, Article 2(b)).
- A differentiation in treatment is not discrimination when the situations are not comparable (EU 2004, Recital (12)). In fact, in such a situation non-differentiation may be discrimination. An example are maternity protection measures that must discriminate between women and men because only women can give birth or breastfeed. Examples include EU (2004, Recital (24)) and EU (2006, Article 15).
- A differentiation in treatment is not discrimination when it is justified by a legitimate aim and the means of achieving that aim are appropriate and necessary ("proportional") (EU 2004, Article 4 (5)). Examples are single-sex sports clubs or shelters for abused women. In specific employment situations, a discrimination-indexed feature may actually be a "genuine occupational requirement". For example, it is legitimate to consider only male applicants when searching for models for men's fashion.

These rules, and therefore also the definitions of which situations are comparable and which are not, and which aims are legitimate and which are not, may change over time. For example, men and women may be argued to be in non-comparable situations when it comes to statistical life expectancy or risk of illness and accidents. Until 2012, Article 5(2) of EU (2004) allowed Member States to "permit proportionate differences in individuals' premiums and benefits [from insurance and related financial services] where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data". On 1st March 2011, the European Court of Justice ruled that Article 5(2) was in breach of the Charter of Fundamental Rights and therefore void, after a transition period lasting until 21st December 2012 (European Court of Justice 2011).

In the employment sector, the legally admissible exclusions of women from certain professions, especially in the police and armed forces, are gradually eroding along with the assumptions that women are "by nature" not suited to them (Pitt 2009). Moreover, the legal provisions of what constitutes illegal discrimination may be quite heterogeneous even across jurisdictions governed by the same principles [concerning insurance, see Schanze (2013) for an overview of pre-2012 European

implementations and Avraham et al. (2013) for an overview of US states' legislations].

Three further aspects are needed to distinguish between the notions of discrimination and related concepts. First, discrimination in a wide sense can involve a merely cognitive making of a distinction, or a making of a distinction in treating people, or a making of a distinction in treating other creatures or things. Discrimination in the narrow and in the legal sense focus on differentiations in treating people. Second, a statistical imbalance in itself is not discrimination—discrimination is a property of a decision or decisions, which may result in statistical imbalances as well as the situation of individuals. As an example, more men than women having jobs in higher management is a statistical imbalance, although it may well be the result of discriminatory decisions. On the other hand, a woman not getting a job just because of her gender is discrimination. Third, discrimination can happen intentionally or unintentionally.

2.2 Classical discrimination-aware data mining (DADM)

In its descriptive role, data mining may *detect* discrimination in a data set, when statistical imbalances originate in earlier decisions. If imbalances result from something else, such as a law of nature, the detected patterns are not discrimination. Establishing the causal reasons of these imbalances of course requires going beyond the mere statistics of data mining. DADM methods are extensions of standard data mining that leverage background knowledge about discrimination-indexed features and their correlation with other features in order to detect discrimination in the narrow sense.

In its prescriptive role, the very point of data mining is to *create* discrimination—in the wider sense: a decision rule by definition makes distinctions based on some features. The basic idea of DADM was to turn this around and use an analysis of its patterns to *prevent creating* discrimination in the narrow sense: If discrimination per se is allowed and desired, but discrimination based on a well-circumscribed set of grounds is forbidden, then data-mining methods must prevent the generation of "bad patterns" or identify them and filter them out.[3] The remaining patterns are by definition "good" ones. Prevention is realised by a number of pre-processing and in-processing methods for DADM, and identification/filtering by a number of post-processing methods. Examples include Hajian and Domingo-Ferrer (2013), Mancuhan and Clifton (2012) (pre-processing), Calders and Verwer (2010), Kamiran et al. (2010, 2012), Kamishima et al. (2012) (in-processing), and Calders and Verwer (2010), Pedreschi et al. (2009), Ruggieri et al. (2010) (post-processing).

As an example, we consider a typical use of data mining: the analysis of old loan data to derive rules for future loan decisions. The descriptive and prescriptive roles of data mining are linked by a set of assumptions: (a) the descriptive analysis revealed imbalances that identify certain features to be predictive of undesirable outcomes (e.g., loan applicants with these properties often default on their loan), (b) existing customers and potential future customers are drawn from the same

[3] "Bad patterns" correspond to, e.g., "α-discriminatory rules" in Pedreschi et al. (2008).

population, and thus (c) decision rules that discriminate against customers with features that have been found to be predictive of undesirable outcomes in step (a) will reduce the occurrence of these undesirable outcomes. We have used this example of loan decisions as the basis for the user studies described in this paper (see Sects. 3.4, 4).

In this view, DADM is therefore but a constraint on step (c), and the reduced utility of forgoing some rules must be outweighed by the (legal or otherwise) need to prevent discrimination in the narrow sense.[4] We therefore call this classical approach to DADM *constraint-oriented*.

Further constraints are imposed on this form of DADM in order to also prevent indirect discrimination such as red-lining. DADM approaches such as those of Calders and Verwer (2010), Hajian and Domingo-Ferrer (2013), and Ruggieri et al. (2010) formalise and take measures against such indirect discrimination.

2.3 The need for exploratory DADM

The constraint-oriented approach to DADM, however, forgoes the advantages inherent in descriptive data mining: the exploration of data that may lead to new insights and new hypotheses to be tested. This is of utmost importance in the field of discrimination too. An exploration of data may lead to insights about new or changing forms of or grounds for discrimination, and it may lead to a pinpointing of (sub-)groups at risk within groups more obviously in danger of discrimination.

One example that is currently being discussed in sociology are the changing challenges that women face in the workplace. Overt discrimination against women appears to have abated relative to the past, thanks in no small measure to past efforts to detect gender discrimination, raise awareness about it, and implement equal-opportunities policies. However, it increasingly appears that *mothers* now suffer from discrimination in the workplace (Fine 2010). This is not only socially relevant, but also a prime example of an emerging pattern that even a typical indirect-discrimination analysis may not notice, since the (not discrimination-indexed) feature "parenthood" is hardly predictive of gender. Such forms of discrimination can only become successful targets for classical DADM if the risks implied by "parenthood" *within* the group with feature "female" have been discovered and a new feature "mother" has been constructed. Note that such feature construction often requires background knowledge and negotiation among stakeholders. For instance, the risks implied by "lack of job experience" (another not discrimination-indexed feature) may be statistically equal to those of parenthood, but are unlikely to be accepted as unjust job-market discrimination. We call such an approach, which focusses on *discovering* features and discrimination, *exploratory DADM*.

An exploratory approach to DADM is also advantageous when it is not clear-cut whether a distinction by some attribute amounts to discrimination in the legal sense or not. Making a feature visible may allow for more open-ended interpretations and evaluations and, importantly, for an awareness of the complexity of the notion of discrimination as such. Constraint-oriented DADM requires a model in which the

[4] See for example Hajian et al. (2011), Kamiran et al. (2010) for measures of utility.

Table 1 Data mining (*DM*), discrimination, and foci of constraint-oriented DADM (*cDADM*) and exploratory DADM (*eDADM*)

	Discrimination (wide sense)	Discrimination (narrow sense, legal sense)
Descriptive DM	Detection	
cDADM		Assumption-based detection
eDADM		Discovery-based detection
Not DADM-supported DM		Detection is possible
Prescriptive DM	Creation	
cDADM		Prevention of creation
eDADM		Feature evaluation/construction
Not DADM-supported DM		Creation is possible

distinction between discrimination and non-discrimination relies on explicit and binary distinctions between legitimate and non-legitimate attributes. However, this may not always be straightforward. First, the temporal and spatial heterogeneity of anti-discrimination legislation needs to be taken into account when, for example, a DADM software is rolled out in a large multinational company. In addition, the modelling of non-comparable situations may require measures that relate to populations[5] or aims[6] as well as their restrictions by legal principles[7]. The visibility of the features may remind the analyst that additional judgment must be applied before a rule is simply discarded as "illegitimate".

The resulting relationships between data mining and discrimination , as described in Sects. 2.1–2.3, are summarised in Table 1. At this high level of abstraction, data mining has similar relationships to discrimination in the narrow and in the legal senses, even if it there will be important differences in practice. We will return to this in the Conclusions.

DCUBE-GUI (Gao and Berendt 2011) is a DADM system that encompasses several of these roles of data mining for discrimination detection and prevention. DCUBE-GUI employs methods from constraint-oriented DADM (more specifically, it builds on rules mined by DCUBE (Ruggieri et al. 2010)) and complements them by risk scores defined on items or item pairs. The analysis of items addresses a descriptive question (people with what features were possibly discriminated against, or simply appear to be at more risk of bad outcomes) as

[5] E.g. the "actuarial factors related to sex" discussed in Sect. 2.1.

[6] E.g. "Differences in treatment may be accepted only if they are justified by a legitimate aim. A legitimate aim may, for example, be the protection of victims of sex-related violence (in cases such as the establishment of single-sex shelters), reasons of privacy and decency (in cases such as the provision of accommodation by a person in a part of that person's home), the promotion of gender equality or of the interests of men or women (for example single-sex voluntary bodies), the freedom of association (in cases of membership of single-sex private clubs), and the organisation of sporting activities (for example single-sex sports events)." (EU 2004, Recital (16)).

[7] E.g. "Any limitation should nevertheless be appropriate and necessary in accordance with the criteria derived from case law of the Court of Justice of the European Communities." (EU 2004, Recital (16))

well as a prescriptive question (which of these features will be applied in decision rules to the detriment of people). The methods for classifier learning from paired instances and for the use of ontologies proposed by Luong (2011) and Luong et al. (2011) open opportunities for such exploration. DCUBE-GUI displays these results in interactive visualisations, thereby inviting users to engage in exploration and sense-making.

3 Use cases and evaluation criteria for DADM

In this section, we investigate how DADM is evaluated today (with a focus on automated evaluations, see Sect. 3.1) and how the requirements for evaluation change when DADM is seen in the larger context of knowledge discovery and in particular as part of decision support. After a general discussion of key issues (Sect. 3.2), we derive conclusions for evaluations of cDADM and eDADM (Sect. 3.3). In Sect. 3.4, we then summarise a first exploratory user study of eDADM and its limitations as a motivation for the experiment to be presented in the subsequent section.

3.1 Automated evaluations and evaluation criteria of DADM

The evaluation of DADM has so far concentrated on the automated analysis of the patterns obtained by the modified algorithms. These evaluations have a simple success criterion: Ideally, all "bad patterns" disappear. In this view of DADM, an effective data-mining method for preventing discrimination applies an agreed-upon definition of bad patterns and guarantees that it either does not find any such patterns or finds all of them and filters them out. An effective system architecture for preventing discrimination employs effective methods and disables possibly found bad patterns.

The resulting success measures of non-existence include counts of successfully sanitised bad patterns, as well as numbers of missed rules and of newly emerging "ghost rules" found in the transformed dataset but not in the original one (Hajian and Domingo-Ferrer 2013). Success can also be measured by reduced discrimination scores (Kamiran et al. 2010). An overview of metrics is given in Hajian and Domingo-Ferrer (2013). Note that agreed-upon definitions of "bad patterns" are still being developed, cf. Pedreschi et al. (2012), Ruggieri et al. (2010). DCUBE (Ruggieri et al. 2010) and LP2DD (Pedreschi et al. 2009) are systems that focus on detecting all assumption-based bad patterns. Systems focussing on making them invisible/ineffective could be modelled on analogous architectures proposed for privacy-protection such as the one proposed by Berendt et al. (2008).[8]

[8] We claim this analogy due to the focus on hiding and sanitising patterns that privacy-preserving and discrimination-aware data mining share. However, using one does not imply the other, and their relation is in general non-trivial (Hajian 2013; Hajian et al. 2012).

These success measures abstract from the concrete use cases of DADM for decision support, but the literature does suggest measures of success in such deployment scenarios, to which we turn next.

3.2 Considerations for the evaluation of data mining for decision support

Viewed simply, a decision-support system is "good" to the extent that it supports "the right" decisions. However, this concept is too vague and maybe not even definable in general. We therefore consider a number of general considerations for evaluating decision-support systems and interactive data mining and then derive specific lessons for DADM from them.

Pertinent methodology comes from design studies and visual data mining (Sedlmair et al. 2012). We follow earlier work that proposes visualisation, interaction, and information as levels of analysis (Marghescu et al. 2004), but focus more strongly on actionability of the information. Actionability is a key concept in the traditional definition of data mining: "Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data", where "useful" or "actionable" means that patterns should potentially lead to some useful action (Fayyad et al. 1996). We will therefore consider measures of the visibility and saliency (through visualisation) of discrimination-related information and measures of the actionability of patterns for application-related decisions.

It is important that evaluations take real decision-making situations into account as well as possible (Perer and Shneiderman 2009; Plaisant 2004), although the difficulties of acquiring actual decision makers and following them in their actual, often long-term professional routines are well-known. The evaluation practice in specific domains such as medical decision support therefore suggests that laboratory studies are useful and necessary as a first step on the way to evaluation in more naturalistic settings (Kaplan 2001). For these reasons, we will investigate in which real decision-making situations various forms of DADM might be useful, for whom and how. We have conducted controlled user studies with non-expert users and placed them in situations requiring decisions.

Finally, when humans decide with decision support from a machine, they often do this under conditions of uncertainty. Even with the help of data and statistics, complete information and full "rationality" cannot be achieved, and they may also not be desired. Rather, humans typically employ a number of *heuristics*, which have been found to lead to typical decision *biases* (Arnott 2006). The design of interactive decision-support systems can address well-known heuristics and biases (Chen and Lee 2003).

A particularly pervasive heuristic is that of *availability*: an outcome will be considered more likely to happen the easier it is to think of it or its examples. Design guidelines for decision-support systems have emphasized the need to address this, usually by making *more* information available through presentation in the digital system. Translated into our setting, we expect an availability heuristic of the following kind: a factor (e.g., a piece of discriminatory information) will be considered more important in a decision situation the easier it is to think of it.

DADM (and related fields such as privacy-preserving data mining) have, interestingly, led to a situation in which two completely different approaches to availability are being proposed: cDADM focusses on making bad patterns *less* available or completely unavailable, whereas eDADM focusses on making them *more* available (or available at all) through various forms of highlighting. In the following, we will explore these two approaches to availability as design choices and in their role of co-determining evaluation choices. We will also ask to what extent the cDADM approach of making discrimination less visible by "hiding" it will indeed make it less cognitively available.

3.3 Use cases and evaluations of DADM decision support

To the extent that discrimination is static and well-defined in terms of a fixed set of discrimination-indexed attributes that decisions must not be based on, and DADM's role is to act as a constraint, we expect its best use case to be a black-box approach. Ideally, the decision-maker should not even get to see the bad patterns (because they might unduly influence her, leading to intentionally or unintentionally discriminatory decisions).

Typical use cases of such systems will involve decision makers as users. An example are employees of a bank who decide on whether to give a loan or not. These may be the original data owners or third parties receiving the data.

The automated-evaluation criteria of non-existence can be directly translated into measures of *invisibility* of bad patterns in decision-making situations. However, one also needs to ask whether this system-given invisibility still creates *actionable* patterns and leads to the correct or desired human decisions. Thus, *decision quality* should be measured as part of actionability. Of course, evaluation also has to integrate appropriate measures of usability.

In the exploratory view of DADM, the *visibility* of patterns and interactive use cases are key—users must be supported in exploring, making sense of, and inspecting bad patterns further, as well as given the possibility of constructing new features for future analysis.

Typical use cases of such systems will involve actors and users who focus on monitoring other decision makers. Examples are societal organisations such as anti-discrimination centres and commissions, or enforcement authorities. Others could be individuals potentially affected by discrimination or their representatives such as lawyers or social workers, judges having to rule on discrimination-related complaints, and last but not least researchers and activists interested in discovering and investigating patterns of discrimination.

An effective data-mining method for preventing discrimination in eDADM applies an agreed-upon definition of bad patterns and guarantees that it finds (or highlights) them. An effective system architecture for preventing discrimination employs effective methods and makes "bad patterns" visible, interactive, and actionable. Evaluation methods must therefore be based on *visibility*, *interactivity*, and *actionability*. Again, *decision quality* should be measured as part of actionability. As in constraint-oriented DADM, system evaluation also has to integrate appropriate measures of usability.

3.4 Can eDADM support non-expert users in exploring items associated with discrimination? A first, exploratory user study

We conducted an exploratory user study to test whether the DCUBE-GUI (Gao and Berendt 2011) interface can support non-expert users in exploring items associated with discrimination. To make the study more engaging and relevant, we embedded the interpretation of DADM results into a fictitious but realistic setting. We asked people to imagine they were social workers giving advice to a client regarding risk factors for a loan. The idea was to have participants recognise the relative risk of different factors and to transform this into a recommendation to the client—to ask for a loan in a way that avoids the most important negative risk factors and, if applicable, take advantage of positive risk factors. Thus, our hypothesis was that the interface supports these steps (comparison of risk factors, identification of important ones, and translation into a correct and useful recommendation), i.e. that it makes the DADM results visible and actionable.

By postulating a scenario, we take a previous definition of top-level item as given (e.g. being female, being a foreign worker) and then investigate how visible problematic second-level items (e.g. being a young foreign worker) become and can lead to action (giving advice to a member of the social worker's community). To limit the complexity of the study and confounding of factors, we restricted the interaction with the tool severely by giving participants screenshots rather than asking them to interact with the tool. This enabled us to focus on measures of visibility and actionability. In addition, we measured basic usability indicators.

In a series of nine scenarios describing the features of a loan applicant and his or her loan request, participants chose a "best recommendation" for the client. The results showed that the highlighting of the relative risk factors by the eDADM tool DCUBE-GUI enabled participants to readily identify negative and positive risk factors and from them to correctly identify recommendations—a sign of high decision quality.

In addition, the answers and the comments indicated that most participants took the task very seriously and thought about the scenarios. The answers and comments also indicated that many people prefer to think about an application scenario of data mining in a more holistic way than only in terms of numbers and risk scores. They took the life context of scenario personnel's age, family, or business into account, and they commented on the ethics of actors' behaviours in the scenario.

The results show that DCUBE-GUI is effective in making the results of DADM visible and actionable. DADM can be presented in ways that make it relevant and interesting to people, help them understand facets of discrimination and draw correct and actionable conclusions from DADM results.

This exploratory study also presented evidence that eDADM is suitable for detecting discrimination, including new forms of it. Still, there are four aspects of DADM usage that were not addressed. (1) This first study only asked people for interpretations of result configurations that were by design quite clear-cut. Also, users were offered decision options, but not asked to motivate their decisions. (2) The study only used one tool. This restricts the interpretation of its results to an evaluation of the effectiveness of eDADM. As a first extension, eDADM and

cDADM should be compared using decision-support interfaces that are as similar and information-equivalent as possible. (3) The study considered only one user role and use case: a social worker whose task is to detect and advise potentially concerned individuals in the face of given discrimination. This spectrum needs to be extended by the users and use cases we have described as characteristic of DADM's role in preventing and monitoring discrimination. (4) The first study was deliberately exploratory and employed only a small sample.

4 How do eDADM and cDADM support decision-making and reasoning in different settings? A large-scale experimental user study

To address the open questions after the first exploratory user study, we conducted a larger study. In this section, we first specify our hypotheses (Sect. 4.1), then give a non-technical overview of the study's method (Sect. 4.2), followed by a detailed description (Sects. 4.3, 4.4). We describe and interpret the results in Sect. 4.5. A discussion of its limitations will be the subject of the general conclusions of this paper.

4.1 Hypotheses

The purpose of the study was to further investigate the role of DADM for the detection and prevention of discrimination. In particular, we were interested in the relative value of eDADM and cDADM for decision quality in different typical settings. These settings are characterised by different foci on discrimination detection and (non-)creation as outlined above. We also wanted to investigate not only the decisions being made, but also the reasoning towards them. This led to the following hypotheses.

The first two hypotheses concern the role of DADM, exploratory and constraint-based, in supporting and motivating decisions.

H1: DADM supports users in making non-discriminatory decisions based on data-mining results, with more accurate results than not DADM-supported data mining.
H2: DADM supports users in motivating their conclusions in non-discriminatory ways with more accurate results than not DADM-supported data mining.

The third and fourth hypotheses concern the differential advantages of cDADM and eDADM for different settings.
H3: For users focussed on making and motivating their decisions in non-discriminatory ways, cDADM supports more accurate and less discriminatory results than eDADM.
H4: For users focussed on monitoring for preventing discriminatory decisions and motivating these conclusions, eDADM supports more accurate results than cDADM.

4.2 Study overview

We created experimental conditions that differed along the dimensions "mining form" and "setting". As mining forms, we chose cDADM, eDADM and, as control

Bank: You work in a bank, and your responsibility is to prepare loan decisions for your manager: Based on an applicant's data, you propose to either grant or deny a specific loan request. The bank's policy is to draw on data analyses of past loan data.

ADA: You work in a citizen-advice / company-watch center, and your responsibility is to prepare decisions for your manager: Based on an applicant's data, you predict likely outcomes. Your manager will use your predictions to derive "alerts" as to which cases to follow up. Various citizens have turned to the center for help: they want a loan from the bank, but are not sure whether they will get it. The center has access to data analyses of past loan data, and it uses these analyses both to advise individuals and to monitor patterns of lending.

Bank, ADA: Thus, you will receive (a) data describing the requested loan and the applicant, and (b) statistical rules that argue for or against granting the loan, given specific data from (a). Based on the information from (a) and the decision support from (b), you will 1. propose a decision for your manager and 2. motivate that decision.

Fig. 1 The overall task descriptions for the bank conditions (*top* and *bottom*) and for the ADA conditions (*middle* and *bottom*)

conditions, non-DADM data mining (DM for short). As settings, we chose a bank and an anti-discrimination agency (ADA), both focussing on the granting of loans. These correspond to the archetypical applications of data mining in decision support: making and monitoring decisions. This results in 3 (mining forms) $\times$ 2 (settings), i.e. 6 experimental conditions. The settings were introduced to participants via instructions about how to use data-mining results for reaching decisions (see Fig. 1) and instructions to avoid discrimination in the process (see Fig. 2).

The 215 participants of our user study, randomly and approximately equally distributed over the 6 conditions, were then asked to consider a series of loan requests. They were given features of the request and the applicant, and provided with decision-supporting rules of a data-mining tool that was fictitious but based on the principles of the mining form. Bank participants were asked to decide whether to grant the loan or not, and to motivate their decision. ADA participants were asked to conclude whether they considered it likely that the loan would be granted or not, and to motivate their conclusion. Examples of the tool and answer choices are shown in Figs. 3 and 4.

We then analysed the decisions as well as the motivations. The results were analysed with a view to testing the hypotheses H1–H4. In addition, these answers and free-form comments were analysed in an exploratory fashion for further insights into how cDADM and eDADM could help against discrimination, and where potential pitfalls lie.

4.3 Method: notes on operationalisation and terminology

We applied some simplifications when operationalizing the constructs in order to (a) test the formalisations of discrimination employed in today's DADM, (b) maximise experimental control, (c) make the tasks feasible for participants, and (d) obtain a first baseline of results.

Note that banks are not allowed to discriminate against applicants based on gender, marital status, nationality, or age. It is important for { **Bank:** the bank | **ADA:** the center to monitor } that decisions that discriminate based on these features not be taken – even if data from the past suggest it.
If any of your answers (the { **Bank:** recommendation to grant or deny the loan | **ADA:** conclusion as to the likelihood of the loan being granted or denied }, or your motivations) need explanation with regard to possible discrimination, please note this in the free-form answer box.

Fig. 2 Instructions against discrimination in the bank resp. ADA conditions

Dabiku is a Kenyan national. She is single and has no children. She has been employed as a manager for the past 10 years. She now asks for a loan of $10,000 for 24 months to set up her own business. She has $100 in her checking account and no other debts. There have been some delays in paying back past loans.

Fig. 3 Example vignette describing the loan request, used in all conditions. Another example is shown in Fig. 4

First, we applied a simplified definition of the "discrimination" we asked participants to avoid: we restricted the specified attributes to four (gender, marital status, nationality, and age), and we declared any discrimination by these attributes as illegitimate, without exceptions. This was done in order to give our non-expert participants a task of manageable difficulty and a clear-cut instruction ("do not discriminate based on these attributes"). The four specific attributes were chosen (a) as typical discrimination-indexed attributes in many jurisdictions (including the European provisions described above and the US-American Equal Credit Opportunity Act ECOA, which applies to most of our participants) and (b) as compatible with a dataset commonly used in DADM (see Sect. 4.4.4). Like its European counterparts, the ECOA previews exceptions to an absolute prohibition to discriminate based on the listed grounds, and a valid identification of whether some decision is legally discriminatory will generally need to involve a legal expert. To avoid this, we gave the simplified instruction.

Second, we wanted to avoid obtaining results confounded by the choice of any specific data mining algorithm. We therefore decided to implement only the key difference between cDADM and eDADM: whether to hide/remove or to highlight discrimination-indexed features in rules.

Future work will be able to build on our results and introduce higher legal as well as computational and interface complexity into our tasks and materials, in particular through exceptions/legitimate grounds for making distinctions based on discrimination-indexed features.

In the materials, the loan applicant and request were described in terms of *features*. The data-mining rules given to participants as a decision basis, as well as the motivations they could select for their decisions, were based on *risk factors* that subsumed features. For example, "age = 37" is a feature, and "age $>$ 30" is a risk factor. We call features, risk factors, motivations and choices *discriminatory* versus *legitimate* (or *non-discriminatory*) depending on whether or not they involve age,

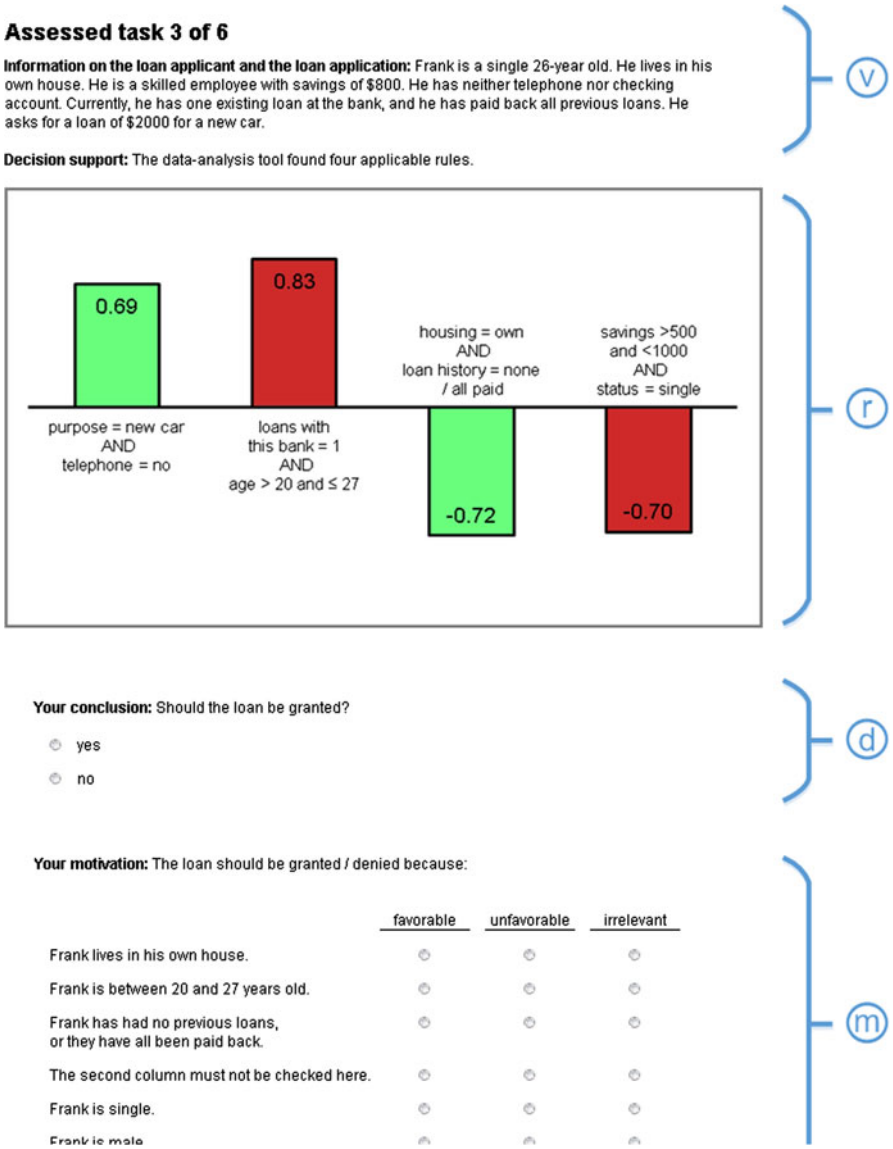


Fig. 4 Example screenshot with *v*ignette, *r*ules for data-mining decision support, *d*ecision, and *m*otivations choice (partial view)

nationality, gender, or marital status. For example, "age $>$ 30" is a discriminatory motivation, and "loan duration $>$ 30" is a legitimate motivation. We call decisions based on legitimate motivations *non-discriminatory* decisions. Note that "discriminatory motivation" is used as a technical term and implies no statements about the psychological motives of the participant.

4.4 Method: details

4.4.1 Participants

In total, 215 US-based participants were recruited over Amazon Mechanical Turk. They received USD 6.00 for full participation and up to USD 1.50 as an additional performance-dependent payoff (bonus). Basic demographics were self-reported in an exit questionnaire (see Sect. 4.5.1).

Sampling through mTurk has attracted some scrutiny with respect to self-selection recently, but it does appear to produce "reliable results consistent with standard decision-making biases" (Goodman et al. 2012). To reduce cultural confounds, we recruited only US participants. We also heeded factors for quality control that have been observed to drastically reduce the occurrence of cheating on mTurk (Eickhoff and de Vries 2013). We included attention-check questions whose cross-evaluation can help identify users who checked answer options randomly. All participants obtained a check score of at least 50 % of the possible maximum. Further analyses of our results gave no indication of cheaters either. Based on these findings, we considered recruitment through mTurk an adequate choice for our study.

4.4.2 Design

The factors *setting* (Bank, ADA as short for anti-discrimination agency) and *mining form* (eDADM, cDADM, DM) were manipulated between subjects.

4.4.3 Procedure

Participants were given a series of scenarios with multiple answer options each. In each scenario, participants ticked exactly one answer corresponding to what they considered the best response for the decision and the relevance of each possible motivation. Three training tasks were presented first after an introductory page with the instructions. The correct answers for the training tasks were shown on the following page, so that participants could check theirs. Six assessed tasks, without information on the correct answers, followed this stage.

An exit questionnaire completed the study. First, we asked for impressions about the task and the tool. Twelve statements were rated on a 7-point Likert scale anchored in "strongly agree" and "strongly disagree". As a simple reliability check, all items came in pairs, with one reverse-coded. The statements build on standard usability questionnaires (Lewis 1995). Subsequently, participants were asked for some basic demographics and personality traits (reciprocity).

Participants were also given the option to comment on the materials, explain their answers, or give any other kind of feedback, by the chance to fill in free-form text fields at the end of each Web page.

All multiple-choice questions (the decisions and motivations, the opinions, and the demographics) had to be filled in; all free-form answers were optional.

4.4.4 Tasks and materials

All tasks had the same basic scenario and overall task, which varied by setting, see Fig. 1. This was given at the beginning. Within this top-level instruction, each participant had to solve three exercise tasks and six assessed tasks.

Each task consisted of four parts. The first was a vignette in which a loan applicant was described briefly, for example by the text shown in Fig. 3. This was identical across all conditions. The second part was the output of a fictitious data-mining tool. In the third part of each task, participants were asked to decide whether to grant the loan request or not (Bank) resp. whether they considered it likely that the request would be granted or not (ADA). Fourth, they judged 12 possible motivations for their decision/conclusion by checking whether these were "favourable", "unfavourable", or "irrelevant" for the decision/conclusion. An example screenshot is shown in Fig. 4.

The tool output consisted of visualisations of decision rules in an intentionally minimalistic way that (a) follows the basic logic of the rule miners that inspired DADM and (b) implements the spirit of the DADM forms and standard data mining. In particular, the tool suggests a "voting" by rules of different strengths for the final decision as in CPAR (Yin and Han 2003), which is also used in DADM (Pedreschi et al. 2008); however it does not perform the last step of calculating the scores that makes the miner decide between two classes ("yes"/"no"). This calculation was left as a task for the user. The tool in its three versions also implements the basic spirit of cDADM (eliminate discriminatory rules), eDADM (highlight discriminatory features in rules), and data mining without DADM support (show all rules, whether they contain discriminatory features or not). Figure 5 shows an example of the three versions.

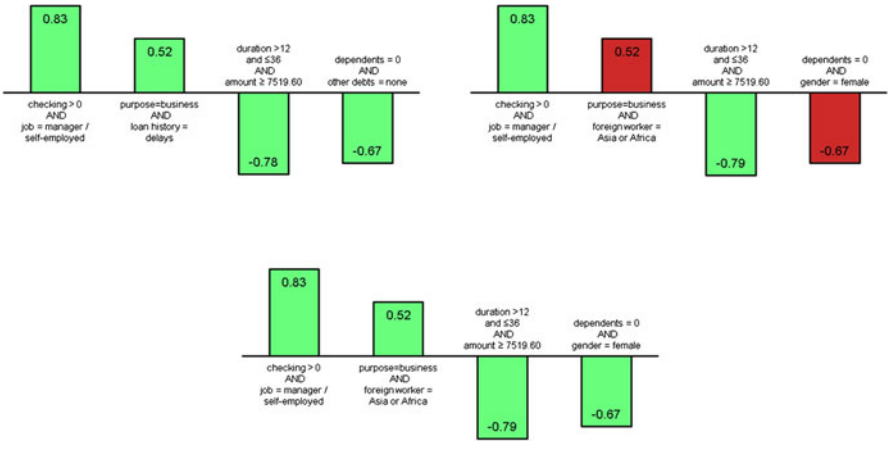


Fig. 5 The tool interfaces for (*top left*) cDADM, (*top right*) eDADM, and (*bottom*) DM. The visualization is identical between cDADM and DM, and the risk factors are identical between DM and eDADM. eDADM highlights rules with discriminatory features in *red* (*second* and *fourth bar* in the example). Identical visualisations were used for the Bank and ADA settings. (Color figure online)

Exercise task (ET) 1 explained the basic logic of rule certainties: Each bar is a rule with one or two risk factors in its premises. All of these must hold in order for the rule to be applied. If the positive risk factors (always above the line) outweigh the negative risk factors, the correct decision is yes, otherwise it is no. ET2 introduced more complex decision settings with several positive and negative rules (two of each). The task explained the basic logic of voting that consists of averaging the certainties of the positive and negative rules, respectively. The materials in ETs 1 and 2 were identical over all conditions. ET2 also gave participants the instruction: "For the following tasks, please remember to answer in line with the policy of your employer of relying on statistically validated results. However, you need not follow the statistical analyses blindly: please exercise judgment where needed." ET3 introduced the topic of discrimination and alerted participants to the need to avoid it, see Fig. 2. As before, feedback was only given on the correctness of the decision.

Assessed tasks (AT) 1 to 6 were like ET3, but without feedback. All assessed tasks were designed equally and with no intentional differences in difficulty.

Risk factors and rule certainties were designed as follows: We created a pool of 17 legitimate attributes and 4 attributes that were explicitly described as discriminatory: nationality, age, gender and marital status. The legitimate attributes comprised further characteristics of the loan applicant (e.g. job status or duration of residence) and of the loan (e.g. loan purpose or duration). These attributes were given a total of 82 values to create features to describe the risk factors in the tasks.[9]

For each task from ET3 to AT6, we randomly chose 3 discriminatory plus 9 legitimate features to describe the applicant and the loan request. The descriptions in ET1 and ET2 had 4 resp. 8 legitimate features. Each feature in any given scenario referred to a different attribute.

From all features describing an applicant, 8 (ET3–AT6) resp. 6 (ET2) or 2 (ET1) were chosen as risk factors for the rules. The risk factors were distributed over the rules to produce 4 rules with 2 risk factors each (ET3–AT6) resp. 2 rules with 1 risk factor each (ET1) or 4 rules with 1, 1, 2, and 2 risk factors (ET2). Distribution was random, except that in both eDADM and DM, 1 positive rule contained 1 discriminatory feature and 1 negative rule contained another discriminatory feature. In cDADM, no rule contained any discriminatory feature. This is shown in Fig. 6.

In ET1 and ET2, the rule certainties implied one correct decision (yes resp. no). In ET3–AT6, the rule certainties were designed such that taking all risk factors and rules into account produced one decision, whereas taking only the legitimate ones into account produced the reverse decision. Thus, the first of these decisions was correct for the cDADM mining form (which had no discriminatory features and thus required that all risk factors and rules be considered), and the reverse was correct for

[9] Our focus was not on analysing any specific true lending data, but on how people deal with data mining results that in reality often are or seem to be non-causal, with correlations often going against common sense and referring to features that act as a positive risk factor in one rule and as a negative risk factor in another one. However, we wanted to create a *possible* loan-related model. We therefore used the attributes of the German Credit Dataset (Newman et al. 1998) as well as their values, and added further values to create a sufficient number of features (for example, we converted the binary "foreign worker" attribute into a multi-valued attribute specifying the country of origin of the loan applicant).

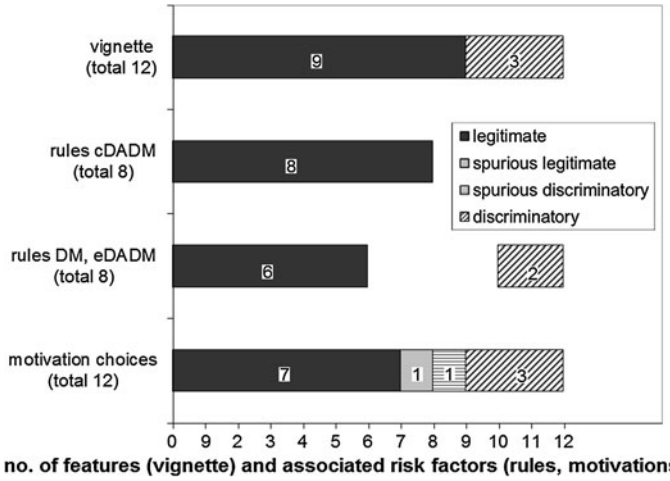


Fig. 6 The construction of features for vignette, rules and motivation choices. The figure gives the numbers of features of the different types. Thus, for example, in eDADM and DM the rules contained 6 legitimate features and 2 discriminatory features taken from the vignette. The motivation choices included all these features, plus 1 extra legitimate and 1 extra discriminatory from the vignette, and 1 extra spurious. (These numbers refer to ET3–AT6; ET1 and ET2 were slightly smaller and simplified)

the eDADM mining form (in which 1 positive and 1 negative rule had to be disregarded to reach a non-discriminatory decision). For 2 of the assessed tasks, the correct cDADM answer was "yes", and thus for 4 of the assessed tasks, the correct eDADM answer was "yes".

For each task, the features mentioned in the vignette, in the rules, and the motivation choices were chosen to ensure that all rules were applicable because they referred to features of the applicant or request. The possible motivations included correct choices (in the vignette, in the rules, and legitimate), irrelevant choices (not in the rules, or in a rule that was irrelevant because its premise also involved a discriminatory feature), discriminatory choices (involving discriminatory features), and spurious choices (not in the vignette). All vignette/rules/motivations designs followed the same schema, illustrated in Fig. 6. All vignette, rule, and spurious choices were random.

Remarks on the unavoidably larger complexity of the ADA task In a sense, the bank setting is more straightforward than the ADA setting. A bank clerk has data and rules (or other data-mining patterns) given by a tool and should make a decision based on this, but not on discriminatory features. An ADA clerk, on the other hand, is faced with an inherently epistemic task in the sense that she has data and patterns and has to make assumptions about somebody else's reasoning and behaviour. These include assumptions about tool access and use, about motivations and decisions, and about one's own role.

Assumptions about tool access and use assumptions could be "I have access to this tool, the bank has and uses the same tool" or "I have access to this tool, the bank has and uses a different tool". Assumptions about motivations and decisions

could be "The bank tries to act ethically" or "The bank does not try to act ethically". One's own role could be perceived more as regulating ("I have to help the bank make ethical decisions") or as monitoring ("I have to detect when unethical decisions were made").

These inherently more complex task aspects are difficult to disentangle and more difficult still to manipulate experimentally. In addition, trying to do so would result in a large increase in the number of experimental conditions, in a situation in which we have no prior empirical knowledge about the workings of DADM in an ADA setting. We therefore decided (a) to use a simple baseline in this first experiment that was as similar as possible to the bank task and designed to draw participants' attention to non-discriminatory decisions, (b) to allow for a certain openness in participants' own interpretation of the setting, and (c) to reflect this in our analysis and interpretation of results.

4.5 Results and discussion

In this section, we describe the results of analysing the decisions and motivations given for the assessed tasks by the 215 participants, divided over the six conditions as shown in Table 2. Additional analyses (Sect. 4.5.6) also investigated exercise-task results. No decision or motivation restricted any other. Also, no indication of dependencies between decisions or between motivations were found in the results.

4.5.1 Participant demographics

Basic demographics were self-reported in an exit questionnaire: 43 % (56 %) of participants reported being female (male). Age ranged from 18 to 69, with a median of 31 years. Among all participants, 12 % reported high school graduate (or equivalent) as their highest grade of schooling, 40 % reported some college (1–4 years, no degree), 38 % a Bachelor's degree, 6 % a Master's degree or a Professional degree, and 2 % "Other".[10] 7 % reported that they "speak a language other than English at home".

A quarter (24 %) reported that they are "dealing with data mining or statistics in [their] job or have done so in the past". 25 % reported that they are "dealing with financial information in [their] job (e.g., banking, insurance, finance industry) or have done so in the past". 13 % reported both. Together, these constituted 36 % of the sample.

Three quarters of participants stated that they had "applied for a loan at least once in [their] life" (73 %, validated by a reverse-coded question), with 50 % of these having at least once been denied a loan. Also, 50 % of all participants reported that they had "experienced discrimination in [their] own life". These proportions mirror those found in our earlier study (Berendt and Preibusch 2012).

[10] The US Census 2012 reports: 85 % (compared to our 98 %) "high school or more", 28 % (compared to our 44 %)"Bachelor's degree or more", 10 % (compared to our 6 %)"advanced degree or more". (http://www.census.gov/compendia/statab/2012/tables/12s0233).

Table 2 Numbers of participants, decisions, and motivations, over all tasks resp. assessed tasks (ATs)

	ADA-cDADM	ADA-DM	ADA-eDADM	Bank-cDADM	Bank-DM	Bank-eDADM
Participants	40	32	32	37	33	41
Decisions (all)	360	288	288	333	297	369
Motivations (all)	3,840	3,072	3,072	3,552	3,144	3,936
Decisions (ATs)	240	192	192	222	198	246
Motivations (ATs)	2,880	2,304	2,304	2,664	2,352	2,952

4.5.2 Decisions [H1]

To analyse decision quality, we investigated the impact of setting and mining form on the number of correct decisions.

We encoded the proportion of "correct decisions" in the assessed tasks as a $2 \times 3 \times 2$ contingency table (2 settings, 3 mining forms, correct/incorrect decisions) and analysed this with log-linear modelling including pairwise comparisons with Bonferroni corrections (Bresnahan and Shapiro 1966). The data are given in Fig. 7. Thus, for example in ADA-cDADM, 240 decisions were made, out of which 184 were correct (as defined in Sect. 4.4.4), which amounts to 76.7 %. Mining form was found to have a clear effect on decision correctness (significant at $\alpha = .01$).[11] Both cDADM and eDADM led to significantly higher proportions of correct decisions than DM, in both settings. No other main effects or interactions were significant.

Participants came to better decisions without taking longer: An investigation of times needed to come to the decisions and give the motivations showed a high variability between participants. On average, the bank setting led to longer response times, with a close-to-significant result in an ANOVA analysis of time-per-task ($p = .06$), and no other significant relationships. However, the DM average was higher for ADA. We were not able to find any other results that correlate with the higher times in the bank conditions.

Taken together, these results support H1: DADM supports users in making non-discriminatory decisions based on data-mining results, with more accurate results than not DADM-supported data mining.

4.5.3 Motivations: overview

But *why* did participants decide or conclude in the ways they did? We analysed the motivations and how they were judged. We partitioned all motivations into x different types and encoded the proportions of these different types in the assessed tasks as a $2 \times 3 \times x$ contingency table, with 2 the number of settings and 3 the number of mining forms. Two different partitionings were designed to take into account the different starting points of the two settings. The first, with $x = 3$ types, is described in Sect. 4.5.4 and the second, with $x = 2$ types, in Sect. 4.5.5. We

[11] All results reported as significant in the following were significant at $\alpha = .01$.

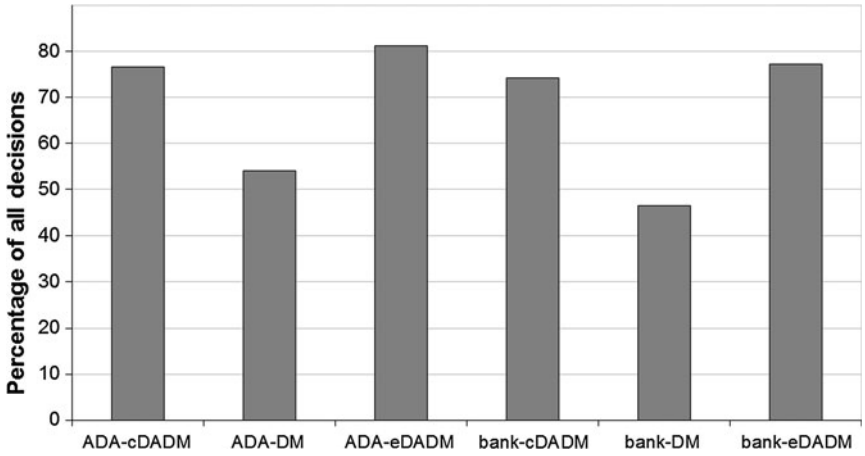


Fig. 7 Percentage of correct decisions by condition

analysed the partitions, including pairwise comparisons, with log-linear modelling, employing Bonferroni corrections.

In addition, we found that discriminatory features were mentioned by participants as relevant for their decisions or conclusions across all conditions. We present and discuss the results of this exploratory analysis in Sect. 4.5.6.

4.5.4 Motivations: the correct specification of legitimate motivations [H2, H3]

The first analysis focusses on the role of DADM for discrimination avoidance. Ideally, DADM would comprehensively ban discriminatory features from the decision discourse and allow decision makers to focus on other reasons for granting or withholding desired treatments. Such avoidance is in line with the major reason for banks to use DADM.

We partitioned the participants' motivations into three groups. (a) *Discriminatory* motivations, as defined in Sect. 4.3, involve nationality, gender, age or marital status. A motivation is discriminatory if the feature was deemed "favourable" or "unfavourable", regardless of whether the applicant has this feature, of whether it is mentioned in a rule, and of whether it is a negative or a positive risk factor. (b) *Avoidance-correct* motivations are features that are legitimate, that the applicant possesses, that are mentioned in one of the task's admissible rules as a positive or negative risk factor, and that the participant correctly identifies as favourable resp. unfavourable. (c) *Avoidance-incorrect* motivations are all others.

The data are shown in Fig. 8. The three-way and all two-way interactions in the contingency table were significant. All pairwise differences except one were significant. Using $>$ to denote a significantly better performance and $\sim$ an insignificant difference, we can summarize:

$$\text{Bank: cDADM} > \text{eDADM} > \text{DM}$$
$$\text{ADA: cDADM} \sim \text{eDADM} > \text{DM}$$

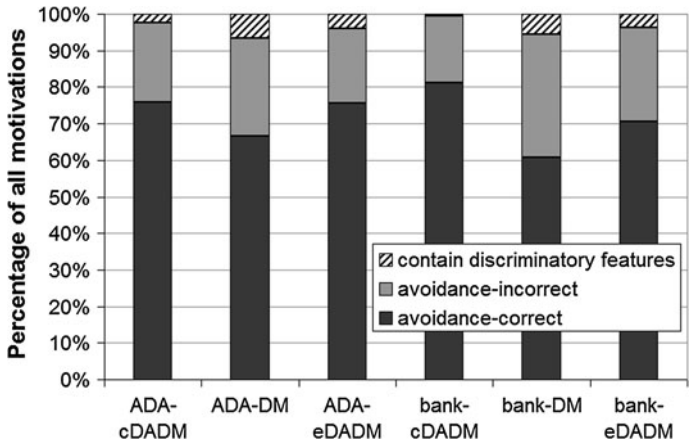


Fig. 8 Detection of correct and non-discriminatory motivations

The bank motivations profited from DADM more and suffered from DM more than the ADA motivations.

Taken together, these results support H2: DADM supports users in motivating their conclusions in non-discriminatory ways with more accurate results than not DADM-supported data mining.

They also support H3: For users focussed on making and motivating their decisions in non-discriminatory ways, cDADM supports more accurate and less discriminatory results than eDADM.

4.5.5 Motivations: the correct detection of discriminatory motivations [H2, H4]

Attention to a discriminatory motivation may mean different things depending on context. For example, some ADA participants indicated, in the free-form comments, that they saw their role as a kind of consultant for the described bank. In such a role, it would be important for them to spot a discriminatory feature/rule *in order to be able to advise, prospectively*, the bank to use other information. An ADA participant may also consider her role to be that of a watchdog who assumes that banks do not necessarily act ethically and therefore needs to spot a discriminatory feature/rule *in order to be able to demonstrate, retrospectively*, that a bank used it. In all such roles, it is key to pay close attention to all rules and risk factors in them.

The second analysis of all motivations therefore focusses on the role of DADM for discrimination detection. Ideally, DADM would comprehensively "spot" discriminatory features in the decision discourse and allow decision makers to focus on the workings of these reasons for granting or withholding desired treatments. Such detection is in line with a major reason for ADAs to use DADM.

We therefore partitioned the motivations slightly differently: (1) *Detection-correct* motivations are all risk factors suggested by the rules, if they are specified with the polarity as indicated in the rule. These comprise all avoidance-correct motivations in the sense of (b) above, and subsets of sets (a) and (c) above.

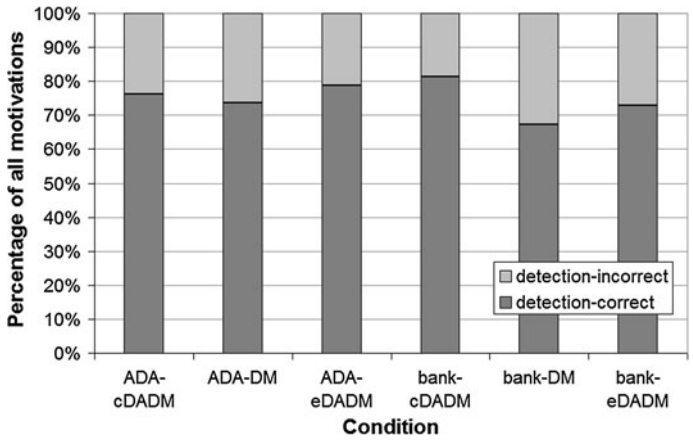


Fig. 9 Detection of given motivations (including discriminatory ones)

(2) *Detection-incorrect* motivations are all others. For cDADM, detection-correct coincides with avoidance-correct, and detection-incorrect covers discriminatory and avoidance-incorrect.

The data are shown in Fig. 9. The three-way and all two-way interactions in the contingency table were significant. All pairwise differences except two were significant. Using the same operators as above and $>\sim$ to denote a near-significantly better performance ($p = .02$), we can summarize:

$$\text{Bank: cDADM} > \text{eDADM} > \text{DM}$$

$$\text{ADA: eDADM} >\sim \text{cDADM} \quad \text{and} \quad \text{eDADM} > \text{DM}$$

Taken together, these results support H2 and also H4: For users focussed on monitoring for preventing discriminatory decisions and motivating these conclusions, eDADM supports more accurate results than cDADM.

4.5.6 Motivations: signs of persisting discrimination?

Although H3 was supported, "less discriminatory" does not mean "not discriminatory". On the contrary, discriminatory motivations were named as relevant (i.e. "favourable" or "unfavourable") across all conditions, including all cDADM conditions in which deciding based on the data mining rules would have involved no discriminatory features, and all bank conditions in which using a discriminatory features clearly violated the bank's obligations. In this section, we report the results of an exploratory analysis of these observations.

Figure 10 shows a further breakdown of the discriminatory motivations. It distinguishes between discriminatory features mentioned in the vignette and in the rules of a task, discriminatory features mentioned only in the vignette, and spurious features, present neither in the vignette nor in the rules. By the construction of the materials (see Fig. 6), vignette-and-rules features constituted half of the possible discriminatory choices in the motivation checklist in DM and eDADM (two of four)

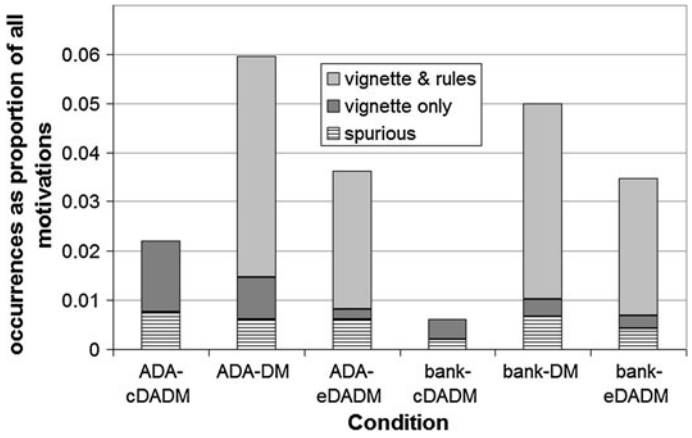


Fig. 10 Discriminatory motivation types

and 0 % in cDADM; vignette-only constituted 25 % resp. 75 %; and spurious features constituted 25 %.

The over-representation of vignette-and-rules features relative to these "prior probabilities" may indicate that motivation specifications were subject to an availability bias. Expressed differently, that the eDADM choice of highlighting rather than hiding a problematic feature may provoke discriminatory thoughts. The presence of spurious features in all conditions may indicate that pre-existing cognitive associations can be activated when judging other people, the typical working of prejudice. The semantics of some spurious discriminatory features suggested this. For example, participants appear to have inferred being married from having children. Alternatively, it may indicate a vulnerability to another cognitive bias, the so-called "Moses illusion" (Erickson and Mattson 1981; Park and Reder 2004): when words and with them thoughts are "put into people's mouth", they are prone to operating with them.[12]

The proportion of discriminatory motivations chosen within the set of all motivations is, fortunately, small. However, the data also indicate that it is persistent: Fig. 11 shows how many participants used at least one discriminatory motivation. Even in bank-cDADM, between 3 and 14 % of participants did this. The figure also suggests that the feedback after ET3 reduced the incidence of such mentions.

Given that we formulated the issue of persisting discrimination as a question rather than as a hypothesis, and that the numbers are relatively small, we do not investigate this subdivision in further statistical detail.

An analysis of the free-form comments revealed possible reasons for checking discriminatory motivations. First, discrimination may be seen only when it is explicitly negative—thus, a rule in which a discriminatory feature is named as a positive risk factor is not considered problematic. In other words, the fact that this

[12] The original observation was that when asked "How many animals of each kind did Moses take on the Ark," most people respond "two," even though they know that it was Noah, not Moses, who took the animals on the Ark.

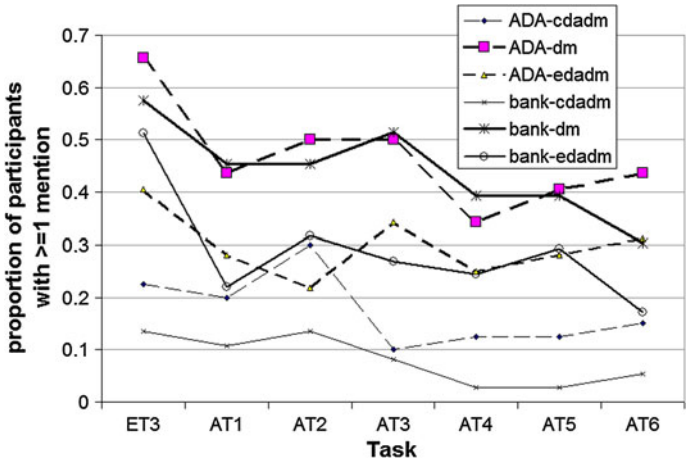


Fig. 11 Participants who mentioned at least one discriminatory motivation

very rule discriminates against people with a different value of the same attribute is not perceived. The data show some evidence of this: 80 % of the discriminatory motivations were rated as "irrelevant" when these features had been mentioned as a negative risk factor, compared to 75 % when they had been mentioned as a positive risk factor. Second, comments indicated a focus on nationality and gender as discriminatory, such that age and marital status were sometimes not identified as problematic. Third, some participants indicated their willingness to "reduce discrimination". One participant remarked: "I dropped the −.67 number a little bit because it included her being a female as a reason". Fourth, background assumptions about loan collateral, job status, and prospects of repayment sometimes obscured the view on discrimination.

Of course, these observations should not be over-interpreted as indicating that any of our participants thought or acted in a sexist, ageist, or in any other way discriminatory fashion. Rather, we want to point out the effects that different data mining tools and the cues given by them may have on the cognitive salience of discriminatory motivations. Even if a tool (such as our cDADM visualization) does not by itself give cues, the environment in which it is used may. For example, a company may internally and/or externally announce that they "are now using a discrimination-safe data-mining tool". Such an announcement, mimicked by the instructions in our experiment, is in itself a possible cue-giver. What follows from cognitive saliency of discriminatory motivations is of course a question for further research.

In sum, even if cDADM's hiding of discriminatory features from data mining *improves* decision making with respect to discrimination, it may not *eliminate* discrimination. Future work should investigate how to reduce the cognitive availability of discriminatory reasoning for decision-making situations like those of our fictitious bank clerk further, and how to reduce the generation of spurious, discriminatory reasoning across all settings.

4.5.7 Opinions on the tool, the task and the participant's own performance

In addition to measuring participants' performance with the tool, we also asked them for usability feedback. Building on a standard instrument (Lewis 1995), participants had to rate twelve statements on a 7-point Likert scale anchored in "strongly agree" and "strongly disagree". They covered the ease of understanding the vignettes, questions and the interface; enjoyment of the task and self-assessed performance at it; as well as intent to reuse the tool for future applications. The items were presented in a randomised order and consisted of six pairs, with a positively and a negatively worded version each. Cronbach's alpha of the overall instrument was $\alpha = 0.90$. The pairwise Pearson correlations between the items and their reverse-coded equivalents were between 0.53 and 0.77, suggesting an overall good reliability.

In general, participants appeared to like the tool, although their feedback was not overly enthusiastic. Of all participants, 62 % agreed or strongly agreed they found the interface easy to understand. 65 % found the questions understandable. More than half of the participants believed they had answered the questions correctly. This self-assessment correlated at $\rho = 0.26$ with their actual performance as the number of correct decisions ($R^2 = 0.07$). There were only weak correlations with the other per-item or overall usability ratings.

No clear picture emerged when we compared the usability ratings across the different experimental conditions. In particular, there is no setting or data mining form that scored systematically better.

4.5.8 Free-form comments

Participants made good use of their chance to comment. Every task had a field for free-form comments, and in addition there was the chance to give general feedback at the end. This led to a maximum of 10 comments per person (based on the data, we aggregated the two general-feedback data items into one). On average, each participant gave 3.3 comments. No clear differences emerged between the settings, but the fewest comments were given in the cDADM conditions, more in the DM conditions, and most in the eDADM conditions. The increase towards eDADM was clearer for ADA than for bank. Averages per condition were: 2.6 (ADA-cDADM), 2.4 (ADA-DM), 4.0 (ADA-eDADM), 2.7 (bank-cDADM), 2.9 (bank-DM), and 4.1 (bank-eDADM).[13]

The comments could be grouped into a number of main content categories, which all occurred in all conditions. (Additional specific content points are described in Sect. 4.5.6.) (a) Some comments just described how arithmetic was applied, such as "The negative risk factors outweigh the positive certainty", some of them enhanced: "Sum of balances is positive after removing discriminatory factors". (b) Many comments indicated that people had been thinking about the scenarios in depth, commenting on the features of the applicant and application and giving (sensible) real-world appraisals of them. They also commented on information that

[13] Due to the exploratory nature of this analysis, we did not test these values for statistical significance.

was *not* mentioned in the rules. Examples of commenting, appraisals, and non-supplied information include "The length of the loan and its small size make it seem acceptable", "Owns a car, so there's collateral", "If it's a business loan, as a lender I'd want to see a business plan before approval". (c) Some comments explicitly described the avoidance of discrimination, such as "Age and nationality must be disregarded, thus the middle two rules are ignored in the analysis" or "If we took into account some of his unfavorable factors we would be discriminating and we don't want that."

Several comments indicated that some participants perceived the study as a test of a new banking tool (and some then commented or complained about the unrealistic rules). Only one explicitly wondered whether this might instead be a "study on how people would react when given the choices presented". Some comments showed visual thinking, i.e. the effectiveness of our interface choices: "Anything that contributed in the RED I marked irrelevant because legally you have to ignore discriminatory attributes." There was a small number of comments on the tool itself, with proposals for interface improvements such as avoiding the need to scroll up and down. 32 participants stated that they had found the attention checks confusing, some indicating worries that they might have given the wrong answers to them, and five more commented on their content otherwise.

Many participants expressed their appreciation of the tasks, for example through "This was unique, interesting, and difficult" or "This was one of the most interesting and enjoyable studies I have done."

5 Conclusions and future work

In this paper, we have investigated how computational methods can help enforce fairness in the knowledge society. Our focus has been on reducing discrimination as a key element of greater societal fairness, and on data mining as one of today's most influential computational methods. In particular, we have presented a conceptual and an empirical analysis of the emerging area of DADM, with a special focus on data mining for decision support.

We have argued for the need to supplement classical, constraint-oriented discrimination-aware data mining by more exploratory forms. We have analysed how constraint-oriented and exploratory forms of DADM are likely to be deployed in practice and what this implies for evaluation. We have summarised the results of a first, exploratory user study, which suggest that DADM can be presented in ways that make it relevant and interesting to people, help them understand facets of discrimination and draw correct and actionable conclusions from DADM results.

In the subsequently described large-scale experimental user study, we have investigated how different forms of DADM can support data mining. We addressed the accuracy and actionability of the conclusions and the reasoning process. The results suggest that both constraint-oriented and exploratory DADM support correct conclusions and reasoning. The results also underline the differential merits of (a) the approach proposed by constraint-oriented DADM to *hide* discriminatory information and thus reduce its cognitive availability and (b) the approach proposed

by exploratory DADM to *highlight* discriminatory information and thus increase users' cognitive awareness. We used decision-making scenarios of a bank and of an anti-discrimination agency as typical examples of two relevant perspectives on whether people are granted loans or not. The results indicate that (a) constraint-oriented DADM can better support users focussed on directly preventing discriminatory decisions, whereas (b) exploratory DADM better supports users focussed on monitoring for preventing that discriminatory decisions are made. We therefore conclude that both forms of DADM complement each other and that appropriate combinations of them will be needed in future real-world tools.

There are of course many aspects of DADM usage that we have not addressed in this study. To conclude, we sketch four aspects as topics of future work.

1. *Tools and study design*: Our studies asked people for interpretations of result configurations that were by design quite clear-cut. Also, users were offered answer options rather than asked to produce answers. In many datasets, less clear-cut relations are likely to hold, and it remains to be seen how interface choices may support or hinder correct interpretations in such cases. It will be particularly interesting to see how the "recall rather than recognition" requirements of open answers will affect cognitive availability and other heuristics and biases.
   Also, participants studied tool output visualisations, but did not interact with the tools. The first reason for this was to make conditions as similar as possible, to reduce cognitive load, and to maximise experimental control. In addition, we believe that this accords well with the current state of the art in DADM, where far more algorithms exist than integrated, interactive deployments of these algorithms in tools. We expect a shift towards more full-fledged tools in the future. It will then be interesting to see how a sequence of exploratory activities and the need to integrate their results in such complex environments will influence visibility and actionability. Extending our methodology of crowdsourcing user-study participants along these lines will be a research challenge that can build on recent work on the evaluation of interactive tools with crowdsourcing (Zuccon et al. 2013).
2. *Notion of discrimination*: As explained and motivated in Sects. 2.1 and 4.3, our study defined the discrimination to be avoided in an intentionally simplified way. The discrimination to be avoided in practice—the one in a legal or even in a sociological sense—is more complex and can often not be reduced to the mandate to avoid differentiating by one or several given features. Future DADM decision-support systems will have to go beyond data mining to be able to deal with decision context, exceptions, and other legally relevant circumscriptions of discrimination, and future DADM research should become a dedicated interdisciplinary area.
3. *Transparency*: eDADM in particular, by its focus on making decision grounds and valuations attached to them visible, can serve as a transparency tool (Gutwirth and De Hert 2006)—an instrument that can make the decision-making of institutions (private or governmental) more understandable. First, it could help make the decisions of monitored institutions (as in the ADA setting)

or of one's own institution (as in the Bank setting) more transparent. Second, it could not only increase understandability for people directly involved in decision-making or in monitoring decision-making, but also for citizens in general. These are the intended beneficiaries of the transparency called for today throughout the world, including the EU and the US, e.g. (European Commission 2012; Federal Trade Commission 2012). The purpose of such transparency tools is to "compel government and private actors to 'good practices' by focusing on the transparency of governmental or private decision-making and action" (Gutwirth and De Hert 2006, p. 9). This can also help achieve more accountability (Alhadeff et al. 2011). To realise this potential, future work on eDADM will need to develop methods that can present data and decision-making to citizens in a usable way and at the same time respect the data-privacy and intellectual-property constraints under which decision-making institutions operate.

eDADM also has the potential to enhance transparency in another sense. Recently, cDADM authors have observed that some patterns of differentiations may be explainable by correlations of discrimination-indexed features with legitimate grounds for differentiation—for example, "no known savings" (Luong 2011, p. 59) as a legitimate ground for rejecting a loan application, or women on average missing specific requirements for a job (Kamiran et al. 2013; Kamishima et al. 2012). The authors have proposed modifications to their algorithms that essentially split an observed pattern of differentiation that appears to be discriminatory into the variance explained by these legitimate grounds and the residual variance that expresses the "real" discrimination by a discrimination-indexed attribute. However, such real-life patterns can also be interpreted in terms of the "intersectionality" of real-life discrimination: the observation that multiple factors of societal disadvantages tend to intersect (Knudsen 2006) (such as specific ethnicities, genders, and ages, low educational level, and poverty). The cDADM approach to "explain away" differentiation may often guard against inappropriate assumptions about decision makers' intentions, but it also effectively hides patterns of intersectionality. In contrast, the eDADM approach can serve to make these very patterns of intersectionality more transparent.

4. *The role of data mining*: We have concentrated on how data mining can contribute to, or help prevent, discrimination by virtue of how patterns are processed and/or presented. However, data mining may also contribute to discrimination in the narrow sense by virtue of its features rather than its patterns.

First, using an attribute at all draws attention to a differentiation that may as well not be made, whereas not storing and/or using an attribute such as nationality would avoid this. This claim is supported by findings from domains as different as peer-reviewing in science and job applications without gender, where the evidence suggests that a decision maker who does not know an attribute's value (the name of the paper's author, the gender of the job applicant) may make choices that are less biased and ultimately lead to better-quality publications or applicant short-lists. On

the other hand, if these features are also unknown to monitoring stakeholders, these may not be able any more to find patterns of indirect discrimination. This might be addressed by sophisticated solutions of differentially disclosed information.

Second, data mining not only uses, but also often produces features. An example are the "profiles" found as patterns in uses such as user/customer modelling. Such profiles are then ascribed as features to new individuals, and this may perpetuate or introduce new discrimination (Berendt 2012). eDADM, by its exploratory nature, can also increase transparency by making such by-products of data mining and new forms of discrimination visible—and thus contribute to more reflection, societal discussion and ultimately better decision making. However, "fairness-aware" computational methods (Kamishima et al. 2012) by themselves cannot ensure social fairness, and they may have side-effects. For example, when insurance tariffs may no longer discriminate by sex, but new sensors and data (of eating habits, sports performance, driving style, etc.) are readily available and their use permitted, the data mining of such data becomes very attractive. Especially when the notion of distributional justice underlying the use of data mining remain stable (for example, premiums based on individual risk factors rather than ability to pay), "eradicating" one form of discrimination may merely shift problems. To the extent that the newly identified desired behaviours indeed are under the control of the individual, new social norms (of eating, movement, and other behaviours) get created and enforced, which can severely limit individual freedoms. To the extent that behaviours are not or only partially under the control of the individual and/or that multiple factors of societal disadvantages intersect, patterns of exclusion will be maintained or shift only marginally. Bringing transparency into *these* patterns is an interesting challenge for fairness-aware data mining—but changing the reality of these patterns also requires legal reasoning and concrete decisions beyond the choice of computational approaches.

Acknowledgements We thank Brendan Van Alsenoy and Albrecht Zimmermann for many inspiring discussions and valuable comments on an earlier version of the paper, and the Flemish Agency for Innovation through Science and Technology (IWT) and the Fonds Wetenschappelijk Onderzoek—Vlaanderen (FWO) for support through the projects SPION (Grant Number 100048) resp. Data Mining for Privacy in Social Networks (Grant Number 65269).

References

Alhadeff J, Van Alsenoy B, Dumortier J (2011) The accountability principle in data protection regulation: origin, development and future directions. Presented at the privacy and accountability 2011 conference, Berlin, 5–6 Apr 2011. http://ssrn.com/abstract=1933731. 11 Oct 2013

Arnott D (2006) Cognitive biases and decision support systems development: a design science approach. Inf Syst J 16(1):55–78

Avraham R, Logue KD, Schwarcz D (2013) Understanding insurance anti-discrimination laws. Technical report. U of Michigan law & econ research paper no. 12-017; U of Michigan public law research paper no. 289; U of Texas Law. Law and econ research paper no. 234; Minnesota legal studies research paper no. 12-45. http://dx.doi.org/10.2139/ssrn.2135800. 20 Aug 2013

Berendt B (2012) More than modelling and hiding: towards a comprehensive view of web mining and privacy. Data Min Knowl Discov 24(3):697–737

Berendt B, Preibusch S (2012) Exploring discrimination: a user-centric evaluation of discrimination-aware data mining. In: Vreeken et al. (2012), pp 344–351

Berendt B, Preibusch S, Teltzrow M (2008) A privacy-protecting business-analytics service for online transactions. Int J Electron Commer 12:115–150

Boston Consulting Group (2012) The value of our digital identity. Liberty global policy series. http://www.lgi.com/PDF/public-policy/The-Value-of-Our-Digital-Identity.pdf. 20 Aug 2013

Bresnahan J, Shapiro M (1966) A general equation and technique for the exact partitioning of chi-square contingency tables. Psychol Bull 66:252–262

Calders T, Verwer S (2010) Three naive Bayes approaches for discrimination-free classification. Data Min Knowl Discov 21(2):277–292

Chen JQ, Lee SM (2003) An exploratory cognitive DSS for strategic decision making. Decis Support Syst 36(2):147–160

Duhigg C (2009) What does your credit-card company know about you? New York Times, 12 May 2009. http://www.nytimes.com/2009/05/17/magazine/17credit-t.html?pagewanted=all&_r=0. 20 Aug 2013

Eickhoff C, de Vries AP (2013) Increasing cheat robustness of crowdsourcing tasks. Inf Retr 16(2): 121–137

Erickson TA, Mattson ME (1981) From words to meaning: a semantic illusion. J Verbal Learn Verbal Behav 20:540–552

EU (2004/2012) Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:373:0037:0043:EN:PDF. 20 Aug 2013

EU (2006) Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:204:0023:0036:EN:PDF. 20 Aug 2013

European Commission (2012) How does the data protection reform strengthen citizens' rights? http://ec.europa.eu/justice/data-protection/document/review2012/factsheets/2_en.pdf. 20 Aug 2013

European Court of Justice (2011) Case C-236/09, Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres. http://curia.europa.eu/juris/liste.jsf?language=en&num=C-236/09. 20 Aug 2013

Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in knowledge discovery and data mining. MIT Press, Cambridge, MA, pp 1–34

Federal Trade Commission (2012) Protecting consumer privacy in an era of rapid change: recommendations for businesses and policymakers. FTC report. http://www.ftc.gov/os/2012/03/120326privacyreport.pdf. 20 Aug 2013

Fine C (2010) Delusions of gender. The real science behind sex differences. Icon Books, London

Gao B, Berendt B (2011) Visual data mining for higher-level patterns: discrimination-aware data mining and beyond. In: Proceedings of the 20th machine learning conference of Belgium and The Netherlands. http://www.benelearn2011.org/. 20 Aug 2013

Goodman J, Cryder C, Cheema A (2012) Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. J Behav Decis Mak 26:213–224

Gutwirth S, De Hert P (2006) Privacy, data protection and law enforcement. Opacity of the individual and transparency of power. In: Claes E, Duff A, Gutwirth S (eds) Privacy and the criminal law. Intersentia, Antwerp, pp 61–104

Hajian S (2013) Simultaneous discrimination prevention and privacy protection in data publishing and mining. PhD thesis, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Tarragona, Catalonia

Hajian S, Domingo-Ferrer J (2013) Direct and indirect discrimination prevention methods. In: Custers B, Calders T, Schermer B, Zarsky TZ (eds) Discrimination and privacy in the information society, studies in applied philosophy, epistemology and rational ethics, vol 3. Springer, Berlin, pp 241–254

Hajian S, Domingo-Ferrer J (2013) A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans Knowl Data Eng 25(7):1445–1459

Hajian S, Domingo-Ferrer J, Martínez-Ballesté A (2011) Discrimination prevention in data mining for intrusion and crime detection. In: IEEE SSCI 2011

Hajian S, Monreale A, Pedreschi D, Domingo-Ferrer J, Giannotti F (2012) Injecting discrimination and privacy awareness into pattern discovery. In: Vreeken et al. (2012), pp 360–369

Heckerman D (2013) From wet to dry: how machine learning and big data are changing the face of biological sciences. http://research.microsoft.com/apps/video/default.aspx?id=189426

Kamiran F, Calders T, Pechenizkiy M (2010) Discrimination aware decision tree learning. In: Proceedings of ICDM'10, pp 869–874

Kamiran F, Karim A, Verwer S, Goudriaan H (2012) Classifying socially sensitive data without discrimination: an analysis of a crime suspect dataset. In: Vreeken et al. (2012), pp 370–377

Kamiran F, Zliobaite I, Calders T (2013) Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowl Inf Syst 35(3):613–644

Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Considerations on fairness-aware data mining. In: Vreeken et al. (2012), pp 378–385

Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Fairness-aware classifier with prejudice remover regularizer. In: ECML/PKDD (2), LNCS, vol 7524, pp 35–50. Springer

Kaplan B (2001) Evaluating informatics applications—clinical decision support systems literature review. Int J Med Inform 64(1):15–37

Knudsen S (2006) Intersectionality—a theoretical inspiration in the analysis of minority cultures and identities in textbooks. In: Caught in the web or lost in the textbook, pp 61–76. http://iartem.no/documents/caught_in_the_web.pdf. 20 Aug 2013

Lewis JR (1995) IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. Int J Hum–Comput Interact 7(1):57–78. http://hcibib.org/perlman/question.cgi. 31 July 2012

Luong BT (2011) Generalized discrimination discovery on semi-structured data supported by ontology. PhD thesis, IMT Institute for Advanced Studies, Lucca, Italy

Luong BT, Ruggieri S, Turini F (2011) k-nn as an implementation of situation testing for discrimination discovery and prevention. In: KDD, pp 502–510. ACM

Mancuhan K, Clifton C (2012) Discriminatory decision policy aware classification. In: Vreeken et al. (2012), pp 386–393

Marghescu D, Rajanen M, Back B (2004) Evaluating the quality of use of visual data-mining tools. In: Proceedings of 11th European conference on IT evaluation, 11–12 Nov 2004, Amsterdam, pp 239–250. Academic Conferences Limited

Microsoft (2012) New York City Police Department and Microsoft partner to bring real-time crime prevention and counterterrorism technology solution to global law enforcement agencies. http://www.microsoft.com/en-us/news/Press/2012/Aug12/08-08NYPDPR.aspx. 20 Aug 2013

Newman DJ, Hettich S, Blake CL, Merz CJ (1998) UCI repository of machine learning databases. GCD at http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29. 20 Aug 2013

Park H, Reder ML (2004) Moses illusion. In: Pohl FR (ed) Cognitive illusions, pp 275–291. Psychology Press, London

Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of KDD'08, pp 560–568. ACM

Pedreschi D, Ruggieri S, Turini F (2009) Integrating induction and deduction for finding evidence of discrimination. In: ICAIL, pp 157–166. ACM

Pedreschi D, Ruggieri S, Turini F (2009) Measuring discrimination in socially-sensitive decision records. In: SDM, pp 581–592

Pedreschi D, Ruggieri S, Turini F (2012) A study of top-k measures for discrimination discovery. In: SAC '12, pp 126–131. ACM, New York, NY, USA

Perer A, Shneiderman B (2009) Integrating statistics and visualization for exploratory power: from long-term case studies to design guidelines. IEEE Comput Graphics Appl 29(3):39–51

Pitt G (2009) Genuine occupational requirements. EC anti-discrimination legislation for legal practitioners, 27–28 Apr 2009, Trier, Germany. http://www.era-comm.eu/oldoku/Adiskri/05_Occupational_requirements/2009_Pitt_EN.pdf. 20 Aug 2013

Plaisant C (2004) The challenge of information visualization evaluation. In: Costabile MF (ed) AVI, pp 109–116. ACM Press, New York

Romei A, Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. Knowl Eng Rev (to appear). doi:10.1017/S0269888913000039

Ruggieri S, Pedreschi D, Turini F (2010) Data mining for discrimination discovery. TKDD ACM Trans Knowl Discov 4(2):1–40

Ruggieri S, Pedreschi D, Turini F (2010) DCUBE: discrimination discovery in databases. In: Proceedings of SIGMOD'10, pp 1127–1130

Schanze E (2013) Injustice by generalization. Notes on the Test-Achats decision of the European Court of Justice. Ger Law J 14(2):423–433

Sedlmair M, Meyer M, Munzner T (2012) Design study methodology: reflections from the trenches and the stacks. IEEE Trans Vis Comput Graphics 18(12):2431–2440

Shearer C (2000) The CRISP-DM model: the new blueprint for data mining. J Data Warehous 5(4):13–22

Sykes JB (ed) (1982) The concise Oxford dictionary, 7th edn. Oxford University Press, Oxford

Vreeken J, Ling C, Zaki MJ, Siebes A, Yu JX, Goethals B, Webb GI, Wu X (eds) (2012) 12th IEEE ICDM workshops, Brussels, Belgium, 10 Dec 2012. IEEE Computer Society

Yin X, Han J (2003) Cpar: classification based on predictive association rules. In: Barbará D, Kamath C (eds) SDM. SIAM, Philadelphia, PA

Zuccon G, Leelanupab T, Whiting S, Yilmaz E, Jose JM, Azzopardi L (2013) Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. Inf Retr 16(2):267–305

6

PAPER 4:

Choice architecture for human-computer interaction

The full paper can be found at http://dx.doi.org/10.1561/1100000028 .

Foundations and Trends® in Human-Computer Interaction
Vol. 7, No. 1–2 (2013) 1–235

DOI: 10.1561/XXXXXXXXXX

Choice Architecture for Human-Computer Interaction

Anthony Jameson
German Research Center for Artificial Intelligence, Germany
jameson@dfki.de

Bettina Berendt
KU Leuven, Belgium
Bettina.Berendt@cs.kuleuven.be

Silvia Gabrielli
CREATE-NET, Italy
silvia.gabrielli@create-net.org

Federica Cena
University of Torino, Italy
cena@di.unito.it

Cristina Gena
University of Torino, Italy
cgena@di.unito.it

Fabiana Vernero
University of Torino, Italy
vernerof@di.unito.it

Katharina Reinecke
University of Michigan, U.S.A.
reinecke@umich.edu

Contents

Abstract

People in human-computer interaction have learned a great deal about how to persuade and influence users of computing technology. They have much less well-founded knowledge about how to help users choose for themselves. It's time to correct this imbalance. A first step is to organize the vast amount of relevant knowledge that has been built up in psychology and related fields in terms of two comprehensive but easy-to-remember models: The Aspect model answers the question "How do people make choices?" by describing six *choice patterns* that choosers apply alternately or in combination, based on Attributes, Social influence, Policies, Experience, Consequences, and Trial and error. The Arcade model answers the question "How can we help people make better choices?" by describing six general high-level *strategies for supporting choice*: Access information and experience, Represent the choice situation, Combine and compute, Advise about processing, Design the domain, and Evaluate on behalf of the chooser. These strategies can be implemented with straightforward interaction design, but for each one there are also specifically relevant technologies. Combining these two models, we can understand virtually all existing and possible approaches to choice support as the application of one or more of the Arcade strategies to one or more of the Aspect choice patterns.

After introducing the idea of choice architecture for human-computer interaction and the key ideas of the Aspect and Arcade models, we discuss each of the Aspect patterns in detail and show how the high-level Arcade strategies can be applied to it to yield specific tactics. We then apply the two models in the domains of online communities and privacy. Most of our examples concern choices *about the use of* computing technology, but the models are equally applicable to everyday choices made *with the help of* computing technology.

now Publishers Inc. *Choice Architecture for Human-Computer Interaction*. Foundations and Trends® in Human-Computer Interaction, vol. 7, no. 1–2, pp. 1–235, 2013.

DOI: 10.1561/XXXXXXXXXX.

1

Introduction

1.1 What Is Choice Architecture for HCI?

If you work in human-computer interaction, you are probably a *choice architect*—even if you have been as unaware of that role as Molière's "bourgeois gentleman" was of having spoken prose all his life.

As Thaler and Sunstein [2008] wrote when introducing the term: "A choice architect has the responsibility for organizing the context in which people make decisions" (p. 3). And users of today's ever-present computing technology are constantly making small choices and large decisions:

1. Sometimes, the main purpose of an interactive system is to help people make a particular type of choice: Think of e-commerce websites and of apps for helping people choose healthy food.

2. Even if the main purpose is different—as with a navigation system that helps you follow a route from one place to another—the user often has choices to make about details—such as which of the several proposed routes to follow. Helping people make these "microchoices" (2.1) better is one (often not obvious) way of enhancing the user experience.

3. Finally, just about any interactive system, regardless of its purpose, requires its users to make some choices about how to operate the system: Which of these two text entry methods should I use to enter text right now? How might I configure this application so as to make it more convenient to use? And might I be better off using some other application instead of this one?

In all of these cases, the fact that the choice is "up to the user" does not release the designers from their responsibility as choice architects to "organize the context" so that users can easily make choices that they will ultimately find satisfactory. But fulfilling this responsibility is easier said than done, if we want to go beyond reliance on designer intuition and familiar design patterns. Good choice architecture for human-computer interaction (HCI) must ultimately be based on a solid understanding of two complex topics:

- The psychology of choice and decision making: How do people go about making choices in their everyday lives, with or without computing technology?

- Strategies and technologies for supporting everyday choice: What are the general ways in which it's possible to help people make better choices; and how can these be applied in the context of—and with the help of—today's interactive computing technology?

This publication aims to equip readers with a coherent understanding of both of these topics, along with an ability to pursue them in more depth by following up on the references. Figure 1.1 gives a preview of the two complementary models that we call the ASPECT and the ARCADE models after their two acronyms: The letters in ASPECT stand for the six *choice patterns* that we introduce to cover the phenomena of everyday choice and decision making. The letters in ARCADE stand for the six high-level *choice support strategies* that we have distilled from previous research and practice.

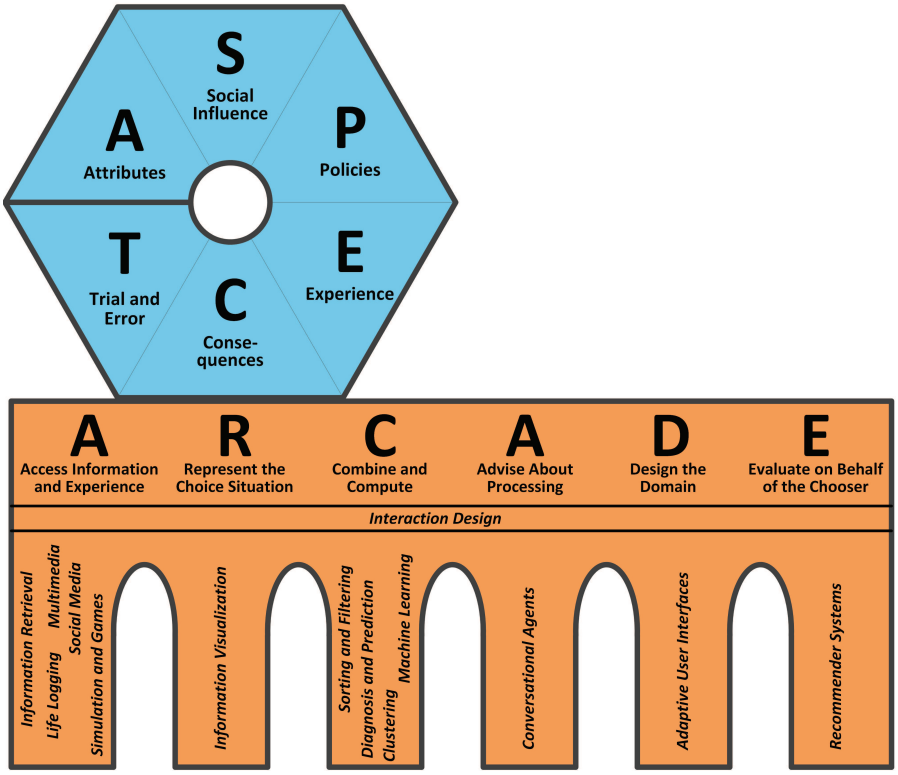


Figure 1.1: High-level overview of the ASPECT and ARCADE models of choice patterns and choice support strategies.

1.2 Hasn't It Already Been Done?

The idea of combining psychology and computing technology to help people make better choices is not new. So why does the HCI field need a new conception of choice architecture? We will explain by first introducing two general conceptual distinctions and then considering in turn several related lines of research and practice.

1.2.1 Preferential vs. Nonpreferential Choice

Many of the "choices" that have received the most attention in the HCI field are *nonpreferential* choices: A user wants to choose the steps (e.g., clicks on particular icons) that are required to achieve a particular goal, such as turning on change-tracking mode in his[1] word processing application. With nonpreferential choices, the question is not what the chooser *prefers* to do but rather what she *has* to do if she wants to achieve a particular goal.

With *preferential choice*—for example, "Shall I turn on change tracking or simply use the commenting functionality to recommend changes to my coauthors?"—a user can prefer one option over another one even though neither one is objectively right or wrong. A preferential choice can be influenced by factors such as the value that the chooser assigns to particular anticipated consequences, the policies the chooser wants to follow, and social expectations that the chooser wants to conform to—a multifaceted set of considerations that will be discussed in connection with the six ASPECT choice patterns.

1.2.2 Persuasion vs. Choice Support

It is also worthwhile to distinguish between two goals that a choice architect can have when attempting to influence a person's choices: *persuasion* versus *choice support*. It is true that neither of these concepts is easy to define crisply and that there are multiple equally reasonable alternative definitions for each concept. Still, there is an important high-level difference between them:

- We will use the term *persuasion* when the goal of the choice architect is to increase the likelihood that the chooser will choose a particular option (e.g., fruit salad instead of cake); or choose an option from some particular class (e.g., fruits and vegetables); or adopt a particular goal (e.g., eat in a more health-conscious way).

[1] To avoid clumsy formulations like "him or her" when using personal pronouns in a generic way, we will alternate between the masculine and feminine forms on an example-by-example basis.

- One possible definition of *choice support* runs as follows: The goal is to help the chooser make the choice in such a way that, from some relevant perspective, the chooser will be satisfied with the choice. One candidate for a "relevant perspective" is: "after learning about the consequences of the choice and taking the time to reflect on all important aspects of it". But other definitions can be argued for. In fact, a first step toward getting better at supporting choice is to understand better what constitutes a "good choice" from the point of view of the chooser (see the discussion in 3.6 below).

These two goals of persuasion and choice support can be pursued simultaneously in various ways. Sometimes, persuasion is used even when the top-level goal is that of choice support. A doctor who tries to persuade a patient to stop smoking presumably believes that the patient will ultimately approve of this choice from some relevant perspective. And in fact maybe the patient has arrived at this conclusion himself and begged the doctor to "persuade" him to perform the specific actions required to stop smoking.

Conversely, even if your top-level goal is to induce a chooser C to choose a particular option O that is in your own interest—for example, the option of buying your software application—adopting choice support as a subgoal can be a good strategy, for either of two reasons:

- You are convinced that C, given high-quality, unbiased choice support, will conclude for herself that O is her best option.

- There are various specific ways of executing O (e.g., various ways of using your software application); and you think that by helping C to choose the specific ways that are best for her, you will increase the likelihood that she will find it attractive to execute O.

Because of these and other interrelationships, techniques for persuasion and choice support can be compared to the black and white keys on the piano (Jameson, 2013): There are some tunes that you can play on just the black keys or on just the white keys; but if you know

how to use all of the keys together, your range of possibilities is vastly increased.

1.2.3 Thaler and Sunstein's Conception of Choice Architecture

Thaler and Sunstein [2008], who coined the term *choice architecture*, present a synthesis of psychological research (chaps. 1–4) that overlaps at many points with the synthesis in our newer ASPECT model, along with six "principles of good choice architecture" (chap. 5), captured with the acronym NUDGES, which suggest how to help people make better choices in everyday life. The remaining 13 chapters of this stimulating and influential book discuss in detail how their principles can be applied in a variety of areas of life, such as personal finance and health.

The relevance of this work for the HCI field is somewhat limited by the fact that Thaler and Sunstein do not devote particular attention to computing technology, either as a means for supporting everyday choice or as a domain in which choices need to be made. Also, as is understandable for a best-selling book, the synthesis of psychological research and the NUDGES principles do not have the clearly articulated structure and explicit grounding in previous literature that is required in a solid foundation for HCI researchers and practitioners. Work that has built on Thaler and Sunstein's conception (e.g., Johnson et al. [2012]) has begun in both of these respects to make the idea of choice architecture more relevant to HCI, but there are still many gaps for the present work to fill.

It is instructive to relate the concept of a *nudge*, which lies at the center of Thaler and Sunstein's conception of choice architecture, to the two conceptual distinctions just introduced above. On close inspection, we can see that the term *nudge* has several different meanings even in these authors' own book:

1. It often refers to a mild form of persuasion intended to bias a person's choice in the direction of a particular option while still being largely compatible with the goal of choice support in that the suggested option seems to be at least reasonably good for the

chooser and in any case the chooser is not compelled to choose it.[2] One of the types of nudge that they suggest (see, e.g., chaps. 5, 6, and 11)—the careful design of default options (cf. 6.2 below)—clearly illustrates this interpretation of the concept of a nudge.

2. Other forms of nudge that they propose—such as structuring complex choices, giving informative feedback, and helping people to "map" information onto concepts that are meaningful for them—can be useful approaches to supporting preferential choice that do not necessarily involve bias toward any particular option. We will be discussing these forms of choice support (along with many others) at many points in the present publication, relating them to the ASPECT and ARCADE models.

3. Finally, several of the forms of nudge can be seen as approaches to supporting *nonpreferential* choice. Under the category "Expect error", the authors present ideas, which will look familiar to readers from the HCI field, about how to help people to avoid doing the objectively wrong thing (e.g., forgetting to attach a document to an email message). Their examples of the nudges in the previous category likewise sometimes concern nonpreferential choice.

The existence of these very different meanings limits the usefulness of the term *nudge* as a way of communicating about tactics for choice support and persuasion. In particular, we may be inclined to agree readily that "people could use a nudge" when we think of the broad meaning that includes any sort of intervention to support or influence choices; but when doing so we can be interpreted as having accepted, in the narrow meaning of the term, a vision of a world in which people's choice processes are constantly being intentionally biased in subtle ways, often without their awareness.

[2] They write: A nudge is "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (p. 6).

1.2.4 Persuasive Technology

When HCI people hear the idea of "helping people make better choices", they often think of persuasive technology: a line of research and practice which was introduced mainly by B. J. Fogg (2003) and which has since become widely represented both in the research literature and in practical systems and interface design methodologies. Like the present publication, Fogg's seminal book systematically combines research from psychology with a framework for making use of the research results in interactive computing technology. Many others have expanded and fleshed out Fogg's framework, and persuasive technology constitutes an important part of a choice architecture for HCI.

A limitation is that persuasive technology focuses squarely on persuasion, as opposed to choice support, as a way of influencing people's choices. It therefore does not provide direct guidance to choice architects who are pursuing the goal of choice support. For this purpose, we need to exploit and organize (in the ASPECT and ARCADE models) a vast amount of literature on choice and choice support that is seldom taken into account in the persuasive technology area.

Paradoxically, our inclusion of concepts and research results that are not oriented toward persuasion may well provide new ideas even to readers who are interested exclusively in persuasion. The reason is that just about every tactic that is designed with the goal of supporting choice can also be (mis)applied in an intentionally biased way (4.7.1). In the present work, we will focus our attention almost entirely on choice support efforts that are not characterized by intentional bias; readers more interested in persuasion will find it easy enough to work out biased versions of any new ideas that they acquire here.

1.2.5 Recommender Systems

A major computing paradigm that can be seen as supporting everyday nonpreferential choice is that of *recommender systems* (see, e.g., Jannach et al., 2011; Ricci et al., 2011). These systems aim to support and influence users' choices concerning products to buy, documents to read, and a variety of other types of item. As we will see in Section 4,

recommender systems essentially implement one of the six ARCADE strategies for choice support, *Evaluate on Behalf of the Chooser*: They typically apply any of a variety of algorithms to predict how satisfied a given chooser would be with particular options. In some cases, an algorithm of this sort can be seen as realizing a variant of one of the six ASPECT choice patterns. For example, some variants of the popular paradigm of *collaborative filtering* (see, e.g., Ekstrand et al., 2011) can be seen as automating a variant of the socially based choice pattern (3.3.4; Section 8), since they make use of information about choices or evaluations made by people who are similar to the current chooser.

1.2.6 Other Contributing Technologies

There are a number of other areas of computer science which, like persuasive technology and recommender systems, contribute techniques that can be used as part of a choice architecture. A number of these are discussed in Section 4 in connection with the ARCADE strategies, which help to explain how they fit into the picture.

1.3 Preview of the Rest of This Publication

Section 2 introduces the several types of choice problem that will yield most of the examples for the present publication. Section 3 offers a compact but broad overview of how people make everyday choices, introducing the ASPECT model. Section 4 introduces the other major part of our conceptual framework, the six high-level ARCADE strategies, giving initial examples of their application and discussing the most important technologies that can be used to realize these strategies. Each of the subsequent six major sections looks at one of the ASPECT choice patterns in more depth, summarizing key ideas from psychological research and discussing how the ARCADE strategies can be applied to support choosing according to the pattern. In the final two main sections, we illustrate how the ASPECT and ARCADE models can help to enhance understanding of choice processes in two important contexts: online communities and privacy, respectively. The final brief section

lists several directions in which the foundation laid in this work can be extended in future work.

12

Choices Concerning Privacy

12.1 Introduction

Whereas in connection with online communities the persuasion perspective has been much more prominent then the choice support perspective, in our second example domain—privacy-related choices—researchers often explicitly state that they are interested in understanding the choices that people make and helping them to make better choices (see, e.g., the comprehensive survey article by Iachello and Hong, 2007).[1] On the other hand, this research literature on the whole makes little reference to the psychology of choice and choice support. Instead, we often see conceptions of choice that are hard to relate to any well-founded psychological concepts. In particular, a lot of the relevant discussion in this area refers to the concept of *privacy preferences*, which is more misleading than helpful, as we will discuss in 12.5.1. One aim of the present section is to show how a clearer understanding of privacy-related choices can be achieved when concepts like this one are replaced with psychologically grounded ones such as those from the ASPECT model.

[1]The principal authors of this chapter are Bettina Berendt and Silvia Gabrielli.

We also wish to call attention to the often underestimated complexity of privacy-related choices relative to most other types of preferential choice in the HCI area. For example, a popular conception is that the quintessential privacy choice is the question of whether to make a photo available to friends or friends-of-friends on a social networking site. We believe that much of the current confusion and concern about poor privacy practices arises from a lack of a principled understanding of privacy-related choices. We therefore dedicate some space to an analysis of this type of choice before showing in subsequent subsections how the ASPECT and ARCADE models can be applied to them.

12.1.1 Three Scenarios for Privacy-Related Choice

We will refer to three scenarios, which can be illustrated as follows with examples:

1. *The social networking site scenario:* A chooser is in possession of a (self-made) photograph of himself, his partner, and their child. The chooser wants to share the photo with some recipient(s) on a social networking site. In the course of this transaction, personal data and information get created and/or transmitted.

2. *The e-commerce scenario:* The chooser wants to purchase some product or service online. The chooser can or has to supply some information *in addition* to the information required for this basic transaction.

3. *The ubiquitous computing scenario:* This scenario covers the main privacy choices of users interacting with ubiquitous and location-enhanced technologies, including sensors (e.g., RFID sensors). The chooser is, for example, an urban traveler who needs to decide whether a particular (group of) people should be allowed to find out her location.

12.1.2 Goals of Choice: What Is Privacy, and What Are "Good" Privacy-Related Choices?

Throughout time and across cultures, people have shown patterns of interacting with others and withdrawing from them. Altman [1976, p. 7] has described these behaviors as *boundary regulation processes*, in

which privacy is "a selective control of access to the self or to one's group". These boundary regulation processes have two essential features. First, privacy involves both closedness and openness: The seeking and the avoiding of social interaction are mutually contingent. Second, the "ideal level" (however one might define that concept) of social interaction or privacy changes with time and context. The ideal level also varies because of individual and cultural differences. These control processes for boundary regulation involve the sharing or disclosing of information vs. the withholding of information, as was investigated in detail in the work of Petronio [2002]. Westin [1967] offers a related notion of *information privacy*: "the claim of individuals, groups or institutions to determine for themselves when, how and to what extent information about them is communicated to others" (p. 7). Definitions such as Westin's (and also their codification in legal regulations on data protection and privacy rights) emphasize that privacy is not only or always about "being let alone" and "disclosing as little data as possible"' (see Berendt [2012] for a detailed discussion).

We will distinguish between three types of "others" and the forms of privacy associated with them:[2]

1. Other people, usually peers. The corresponding form of privacy is called *social privacy*.

2. Commercial entities or other institutions, which raise issues of *institutional privacy*.

3. Governmental entities, which raise issues of *protection from surveillance*.

We define a *privacy-related choice* of a chooser as any choice that results in the granting of access to data about a person to one or more (potential) *knowers*—or withholding access from them. The person whom the data concerns is usually the chooser himself, but it can also be someone else.

[2]These classifications have been discussed by many authors and with different names; we use the terminology introduced in Raynes-Goldie [2012] and Phillips [2004]. *Information privacy* is often associated with "others" of types 1 and 2.

The central privacy goals, in the sense of a *good outcome* for the person (cf. 3.6), are to achieve the level and/or form of privacy desired by the person. As with other choices, not only considerations of outcome play a role.[3]

Design choices that have privacy implications are determined not only by end users' goals but also by the goals of site operators. Often, the underlying business model rests on users paying for a free service with their personal data. Our examples will illustrate support for both types of goals.

12.1.3 A Key Characteristic of Privacy-Related Choices: Multiple Interrelated Choices in Several Dimensions

One key characteristic and challenge of privacy-related choices is that they rarely come alone. Instead, a privacy-related choice is often bundled with another choice, often one that is more important to the chooser. Hence a privacy-related choice will often receive less attention than it deserves simply because of limits on attention and time—or it may even be neglected completely. A narrow focus on only one of the bundled choices can also be encouraged by site operators who are mainly interested in obtaining data, as well as by cultural factors.

The first form of bundling concentrates attention on what may be called a *main choice*, which makes at least part of the privacy-related choice a *secondary choice*:

- The main choice in the e-commerce scenario is how to execute a commercial transaction effectively: The user wants to buy a winter jacket, not make decisions about his privacy.
- The main choice in the ubiquitous computing scenario may be how to reach your destination quickly and safely.

[3]Even this simple definition is faced with a number of challenges. The person's desires may be unknown or unclear, and they may change over time. If the person is not the chooser herself, then the person's desires may conflict with the chooser's wishes. Other people may decide which levels are applicable for the chooser/person, for example when the chooser is below the age of consent or where privacy is regarded as a public good. In the following discussion, we will abstract away from these additional complications.

- The main choice in the social networking site scenario is a choice related to social privacy (e.g., how to show a photo to a particular relative and whom else to show it to). The choices related to institutional privacy—what information should be shared with the site or provider—are often not even perceived.

In all three scenarios, surveillance (i.e., the sharing of information with a governmental entity) generally remains invisible. And although the outcomes of the privacy-related secondary choices may be highly relevant for the chooser, they are often invisible and drawn out over time, and they may be caused by people other than the chooser (see, e.g., Gürses and Diaz, 2013).

A second form of bundling is that one person's privacy-related choices constrain other people's privacy and privacy-related choices, now or in the future:

- In the example given for the social networking site scenario, publishing the photo also discloses information about the other two persons depicted. Interface options for letting the affected persons co-decide are not straightforward and rarely available; and some persons (such as the child in the photo) may not even be able to voice or enforce their decisions.

- Norms of social reciprocity and ease of handling strongly suggest that reactions to a privacy-relevant action should be delivered over the same channel (e.g., social networking site)—which in effect requires people to register and communicate on the same platform.

- Professional norms or concrete activity decisions (such as a decision to teach a class using Google Hangout) can also enforce membership in and activity on certain platforms.

- Choices in e-commerce scenarios may also be constrained by the simple nonexistence of privacy-friendly providers or options, by an excessive price exacted for protecting one's privacy, or by many other factors that make reality differ from an ideal market in which demand for a service can always be met by supply and

no participant is more powerful than any other in determining prices.

- Some constraints on privacy-related choices arise from the joint workings of individual and collective behaviors and technical factors. For example, in order for a system to generate real-time information for ride sharing, multiple parties have to provide location-based data. Another example is the safe encryption of messages with public-key cryptography: This practice requires all participants in an exchange to generate, manage, exchange, and use keys.
- All of these behaviors can also give rise to social examples and expectations (see Section 8) and to habits (7.3) that in turn influence future privacy-related choices in the direction of increased disclosure.

12.1.4 How to Support Privacy-Related Choice? "Nudges" Vs. Awareness Support

In this section on privacy-related choices, we will describe a number of choice support tactics, some embodied in entire tools for supporting privacy-related choices, with reference to the ASPECT and ARCADE models. Many of these tactics can be supported by two categories of tool that currently constitute foci of development and investigation: *privacy nudges* and *data-based privacy awareness tools*. Since tools in both of these categories can support all of the ASPECT choice patterns, we will introduce the basic distinction between them here.

Privacy Nudges

The concept of a *privacy nudge* is nicely illustrated by the three examples presented by Wang et al. [2013]. Although the term *nudge* is commonly associated with influencing choices in a particular direction (cf. the discussion of this concept in 1.2.2), these interface elements can be interpreted naturally within the ARCADE model as forms of choice support. The context is one in which a user of FACEBOOK is about to post some content.

The *picture nudge* shows the user five randomly chosen pictures of members of the set of people who will see the post. This method instantiates the strategy *Represent the Choice Situation* by augmenting an abstract representation of the set of recipients (e.g., "anyone on the internet") with a more concrete representation in terms of particular individuals. This type of representation presumably encourages thinking about how the various possible viewers of the post will respond to it.

If the five pictures presented were not selected at random by the system but rather chosen so as to be representative of the set of people who were actually likely to view the post, this nudge would also instantiate the strategies *Access Information and Experience* and *Combine and Compute*.

The *sentiment nudge* of Wang et al. [2013] alerts the user if the content of the post seems likely to be perceived as "negative", as determined by analysis of its words by a sentiment analysis module. This nudge both evokes the consequence-based pattern and applies the strategy *Combine and Compute* to help the chooser predict a particular type of consequence of her action.

Finally, the *timer nudge* explicitly introduces a delay of 10 seconds after the user has submitted the post, during which he can cancel the post. This method is similar to simply advising the user to "think for 10 seconds before sending off your post", the difference being that the user is almost forced to take the advice (since he has little else to do during the 10 seconds). Hence the nudge implicitly instantiates the strategy *Advise About Processing*. It also illustrates once again (cf. 4.4) that in interactive systems, as opposed to human-human dialog, procedural "advice" often takes the form not of verbal advice but (also) of interaction design that makes it especially convenient or even necessary to take the advice.

Data-Based Privacy Awareness Tools

Data-based awareness tools (see, e.g., Gao and Berendt, 2011; Berendt et al., 2012; Wang et al., 2013) are software tools that apply the strategy *Access Information and Experience*, often in conjunction

with *Combine and Compute* and *Represent the Choice Situation*, in any of various ways to make users aware of potential privacy threats. For example, in support of the consequence-based pattern, the "awareness" can concern the need to make a choice, the assessment of the current situation, and the possible consequences of particular actions (see 12.2 below). In support of the trial-and-error-based pattern, users are made more aware of the consequences of actions once they have performed them (12.6). Data-based awareness tools may—but need not—include privacy nudges.

After this introductory discussion, we will consider each of the ASPECT choice patterns in turn (except for the experience-based pattern), in each case looking at examples of how the concepts associated with that pattern apply to some types of privacy-related choice.

12.2 Consequence-Based Choices About Privacy

It is natural for people to apply the consequence-based pattern to privacy choices, since poor choices can often lead to unfortunate consequences. But this choice pattern raises special challenges for several of the steps in this pattern (see Table 6.1).

12.2.1 Situation Assessment and Recognition of Choice Opportunities

In an ideal world, users would always know when some sort of privacy-related choice was called for and what the relevant features of the current situation were (6.2, 6.3).

In some cases, interface elements indicate straightforwardly where a privacy-related choice does or does not have to be made, as in the following examples, which instantiate the strategies *Access Information and Experience* and *Represent the Choice Situation*:

- Checkboxes for opting out of or into the transfer of information to particular third parties alert the user to a need to choose who will or will not receive particular data.

- A red asterisk in front of a mandatory data field indicates that there is no choice about whether to provide this information.
- The conspicuous provision of telephone contact numbers on the screen signals to the user that she can choose an alternative method of transferring information to the institution in question.

It would in general be infeasible, however, to alert users in a conspicuous way about every privacy-related choice opportunity—even aside from the fact that organizations and individuals who are interested in acquiring personal data often have no interest in increasing users' situation awareness.

For example, in principle, whenever a user visits a website that is going to make use of his personal data in one or more ways, the user might want to choose whether to allow those uses of the data. But often users have no idea what use (if any) is going to be made of their data, and they can hardly take the trouble to find out every time they visit a website. One well-known approach to this problem was the Privacy Bird (Cranor et al., 2006, discussed in 12.5 below), which automatically checked for possibly undesirable consequences and alerted the user to cases where an explicit choice appeared to be worthwhile.

As is discussed by Iachello and Hong [2007, p. 55–57], users of web browsers regularly encounter cues that are provided with the goal of improving situation assessment, such as "lock" icons and warnings that data may be intercepted by third parties. One problem with such cues, which is typical of attempts to support situation assessment via the strategies *Access Information and Experience* and *Represent the Choice Situation*, is that users have only a limited capacity to attend to them, especially since the cues in general do not concern the user's primary task. Even if such a cue is noticed, the user may not be able to make any of the inferences that situation assessment is supposed to support, concerning what will happen if the chooser takes no action and what the consequences of particular actions will be—because of a problem that is mentioned repeatedly in this section: the incomplete mental models that most users have of privacy-relevant factors. For example, if you are informed that data that you enter into a web form might be intercepted by a third party, you probably have little idea of how

likely this event is to occur, who might do the intercepting, or what use they might make of the data. Warnings of this sort are often best seen not as helpful applications of choice support tactics but rather as steps taken to protect a stakeholder from criticism or legal challenges ("Don't blame us; we warned you!").

In the ubiquitous computing scenario, we have the additional challenge that it tends to be more difficult than with graphical user interfaces to provide useful cues to support situation assessment (see, e.g., Nguyen and Mynatt, 2002; Iachello and Hong, 2007, pp. 55–57).[4] In connection with older devices such as audio recorders, warnings such as flashing red lights and regularly spaced beeps have long been used to instantiate the strategy *Access Information and Experience* to support situation assessment. A current challenge is to find analogous methods to provide adequate awareness of the much larger range of sensors that now exist even in an ordinary smartphone (including, e.g., GPS, motion and orientation sensors, and video cameras). The Privacy Mirrors framework of Nguyen and Mynatt [2002] takes into account the broad range of things that users of ubiquitous computing technology in principle ought to be aware of: not only the technical properties of the systems that they are dealing with (cf. Edwards et al., 2001) but also relevant aspects of the physical and social environments.

Many of the specific methods that have been developed for increasing users' situation awareness can also serve to provide feedback in the context of the trial-and-error-based pattern (see 10.5 above and 12.6 below.)

12.2.2 Deciding When and Where to Make a Choice

The question of when a particular choice is best made (6.4) has been discussed especially often in the privacy domain. One key question is whether users can be expected to make configuration choices at one point in time that will relieve them of the need to make a large number of specific choices later on. For example, Lederer et al. [2004, p. 447–448] argue that "Emphasizing Configuration Over Action" is one of

[4] Bellotti et al. [2002] discuss related issues with "sensing systems" on a more general level.

the typical pitfalls associated with the design of interactive systems that have personal privacy implications. Since having users choose privacy settings is essentially a matter of asking them to select a policy that will be implemented (semi)automatically by the system, this issue is discussed below in the subsection on policy-based privacy choices (12.5).

12.2.3 Making an Appropriate Set of Options Available

One source of difficulty in choosing is when an unnecessarily restricted set of options is available that forces the chooser to deal with negative consequences or difficult tradeoffs that would not arise if a more suitable option were available. An illustration of this point is provided by the efforts made within a ubiquitous computing scenario by Iachello et al. [2005] to determine empirically what set of options would be found suitable by users of an application that enables people to disclose their locations to each other. Options that were considered by the study participants to be useful to have included (a) sending a vague, evasive response such as "I am busy"; and (b) sending an inaccurate indication of the user's location (e.g., "on the way home" vs. "still at the office").

On a more general level, Lederer et al. [2004, p. 448] argued that failing to provide "an obvious, top-level mechanism for halting and resuming disclosure" is another of the pitfalls in the design of privacy-relevant systems. They noted as an example that, at the time of their writing, there was typically no simple way for the user of a web browser to block *all* cookies during part of her browsing session. In the intervening years, many web browsers have filled this gap in the set of available options by introducing a "Private Browsing" mode that can be turned on and off easily.

Judiciously expanding the set of options in this way is an application of the strategy *Design the Domain*, which can be favorably compared to the alternative of using other ARCADE strategies to help users to choose among less desirable options.

12.2.4 Holistic Representations of Privacy Choices

One reason for the difficulty of privacy-related choices is that it is often unclear to choosers what the consequences of the various options will be. As we saw in Section 6, there can be many reasons for difficulty in anticipating consequences. One difficulty that seems relatively characteristic of privacy choices in current web-based and ubiquitous systems is due to the way in which options are represented to users—that is, to a problematic application of the strategy *Represent the Choice Situation* (cf. the discussion of framing in 6.7.2): Options are often represented to users in what may be called a *holistic* way that makes it relatively hard to perceive individual consequences clearly. After discussing this type of representation, we will argue for the benefits of more specific representations of individual consequences.

A privacy-related choice situation is often represented by the system's designers as one in which it is desirable to disclose information. This representation can induce users to disclose more information than they otherwise would. In these cases, the strategy *Represent the Choice Situation* is being used with a persuasive intent (cf. 1.2.2) to push the user toward an option that is desirable from the point of view of those who run the system.

One frequent representation is that of *information disclosure as part of an exchange*. This representation draws on real-life characteristics of information disclosure and privacy, especially in e-commerce scenarios. In many commercial and other public transactions, a participant needs to disclose private—or even sensitive—data in order to enable the transaction to occur: In the clothing store, you need to state your size in order to obtain clothes that fit; in the doctor's office, you need to talk about your ailments in order to receive diagnosis and treatment. Many online stores and services promise personalization advantages in return for data. The *perception* that personalization is occurring can lead people to disclose more information about themselves, even if everyone is in fact being given the same recommendations (Kobsa and Teltzrow [2004]). The assumption that questions are asked because they are relevant in the given context may lead people to answer even questions

that would be considered illegitimate outside of the context suggested by the representation (Berendt et al., 2005; Berendt, 2009).

In the social-networking-site or ubiquitous computing scenarios, options that involve information disclosure are often described as involving an exchange of data from the user for a free service from a provider—though the relevant formulations are generally to be found in relatively hard-to-reach documents such as terms-and-conditions pages.

By contrast, the fact that data are being exchanged among peers is made salient by the interface design: Interactive elements for reacting to a new piece of content are placed right next to that content. The fact that such exchanges also feed data to the service provider remains in the background. FACEBOOK typically frames choices made when using the social-networking site as choices about communication with peers and empowering internet users.

Outside of the HCI context, similar framing has been used in the "interfaces" of questionnaires about surveillance choices. For example, in a large-scale survey after the 2013 leaking of details about the PRISM surveillance programme in the U.S., Pew Research[5] asked questions such as "Should the government be able to monitor everyone's email to prevent possible terrorism?" and obtained widespread acceptance.

Another type of representation that may have a persuasive intent involves representing multiple privacy-related choices as if they were basically the same—encouraging broad choice bracketing (9.2)—even if the choices differ quite a bit in terms of their consequences for the user. Figure 12.1A shows the beginning of an explanation of choices about providing location information in iOS 7. Note that the choices are framed as whether to provide location information in order to benefit from a service offered via the iPhone. It is only by reading considerably further that the user can learn that some of the services in question benefit not the individual user who is providing the data but rather users in general—for example, by contributing to a crowdsourced road traffic database. Similarly, in the screen that lists system services (Figure 12.1B), there is no visual distinction between services of the first type (such as "Setting Time Zone") and those of the second type

[5] *Public Says Investigate Terrorism, Even If It Intrudes on Privacy*, 10 June 2013.

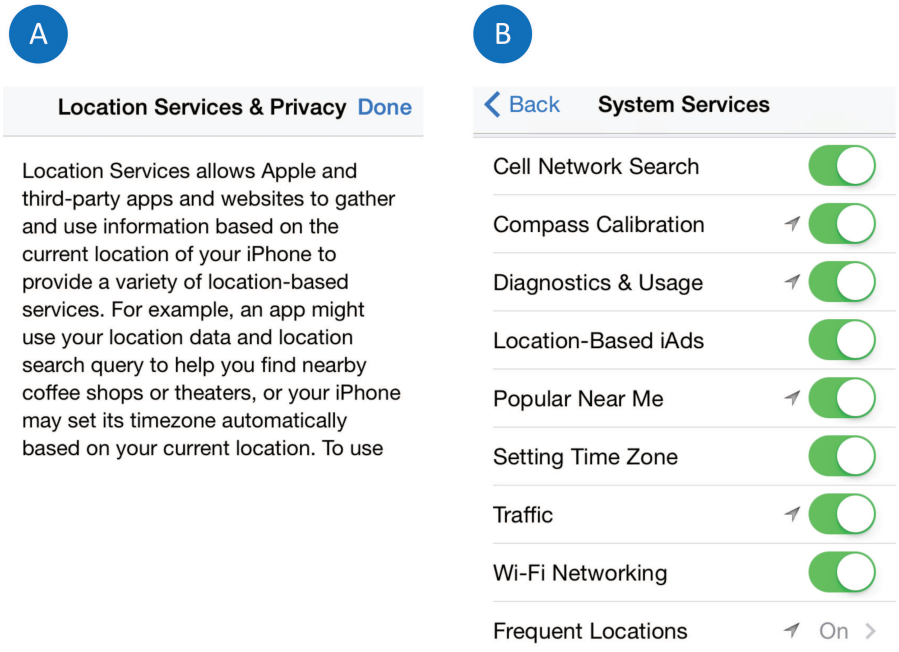


Figure 12.1: Two screens from Apple's iOS 7 that illustrate how significantly different privacy-related choices can be bundled together in a manner that encourages the chooser to deal with them in the same way.

(e.g., "Traffic"). Finally, there is a switch for turning off *all* location services but no such switch for turning off only services of the second type. If we assume that the distinction between the two types of system service is decision-relevant, a natural form of choice support is to make the difference easily recognizable (*Represent the Choice Situation*) and to provide controls for conveniently handling the two types of service differently (*Design the Domain*).

12.2.5 Challenges and Approaches for Predicting Specific Consequences of Privacy-Related Choices

To put it simply, the consequences of privacy-related choices will be that someone will know something and/or that actions affecting the chooser or others will occur, both of which may inflict harm or bring

benefits (Berendt, 2012). Privacy-related choices face just about all of the general challenges (discussed in Section 6) raised by the prediction and evaluation of possibly uncertain future consequences (e.g., hyperbolic time discounting, as was argued by Acquisti and Grossklags [2004]). They also face a number of specific challenges, most notably the general impossibility of undoing choices (12.6) and the often unpredictable consequences of data processing.

A piece of data that is disclosed does not only get recorded and retrieved; it also gets *processed*. It is hard if not impossible to predict what knowledge (and actions) a piece of data—or even the fact that a piece of data is missing—can contribute to after it has undergone (perhaps repeated) linking to other information and processing such as data mining.[6] Cumulative risk assessment—understanding how small pieces of data disclosed over time can be combined to contribute to a profile of a person—is even more difficult.

Choices in ubiquitous computing scenarios face a number of specific challenges related to the fact that the ultimate consequences of particular choices tend to be hard to predict without the benefit of specialized knowledge or experience. In a study of people's responses to five methods for obfuscating GPS trace data collected from them, Brush et al. [2010, p. 7] found that their study participants largely understood the basic operation of the obfuscation methods—but that they had a hard time understanding what the consequences would be of applying a given obfuscation method to data collected over a period of time. For example, participants considered the *subsampling* method of systematically leaving gaps in the collection of data to be basically acceptable, evidently not noticing that its application for an entire day would normally reveal the user's home location—a consequence which, in other contexts, they judged to be unacceptable.

Nguyen et al. [2008, p. 189] asked a broad variety of shoppers in U.S. shopping malls about the possible dangers of six tracking and recording technologies: credit cards, store loyalty cards, electronic toll collection systems, web server records, store video cameras, and RFID

[6] For a seminal HCI perspective on this issue, see Dix [1990]; for a recent overview of challenges and technical solution approaches, see Berendt [2012].

sensors. Whereas respondents showed a clear understanding of the *benefits* of these technologies, "they had difficulties articulating possible costs or threats of these technologies" (p. 186). For example, though they could easily enough anticipate a straightforward possible consequence of divulging their credit card information—that someone might use their data to make an unauthorized purchase—they did not mention the consequences of allowing organizations to build long-term records of their purchases.

In sum, the interrelatedness of privacy-related choices and the invisibilities, indirections, delays, and nondeterministic consequences of the multiple actions that stakeholders perform with personal information (see also 12.1.3) make it hard for a user to predict what may happen as a result of any given privacy-related choice: A very wide or even unknowable range of knowing and action consequences could result from a privacy-related choice. On the other hand, alerting choosers to this wide range of consequences and the difficulties of predicting them could lead to information overload and also fear. This fear, in turn, could lead to inactivity or otherwise ineffective choice. Support approaches therefore tend to focus on specific consequences.

One general difficulty with anticipating the consequences of configuration choices is that a given configuration option can have a large number of consequences for different situations (e.g., adopting a high velocity for a mouse can work out differently from one type of task to the next). One interesting approach to this difficulty, which can be realized with some types of configuration problem, is to give the chooser an overview of the consequences of a given configuration for a wide variety of situations. PViz (Mazzia et al., 2012) helps a user of Facebook to answer questions about the visibility of elements of his profile, such as "With my current privacy settings, which people will be able to see my cell phone number?" PViz generalizes the idea realized in Facebook's *Audience View*, which enables a user to see how her profile will look to a particular individual. To provide an overview of visibility for all possible viewers, PViz does some automatic clustering of people and labeling of clusters, as well as graphical visualization that enables the user to see which proportion of members of each cluster will be able

Figure 12.2: Facebook audience selector tool to alert and support *who* choices about possible recipients when posting information on social media.

to see the profile element in question. It therefore instantiates all of the first three ARCADE strategies. The results of user studies can be interpreted broadly as showing the promise of this approach for helping people anticipate the consequences of policy decisions.

12.3 Attribute-Based Choices About Privacy

Privacy-related choices are occasionally represented to users in a way that encourages attribute-based choice.

A frequent type of choice faced by members of social networking sites is the social-privacy choice of whom (among the other members and the public at large) to share some particular content with. The persons in question may be identified in different ways, such as by their names and/or by their attributes. The basic problem being tackled is the difficulty that users have in perceiving and managing the intended and actual "audiences" of posts given their typically long and undifferentiated sets of contacts.

The answer to this problem is sought in various forms of *access control* that have been derived from models established in the security field, as described in the survey by Sayaf and Clarke [2013]. For example, in FACEBOOK today, users are by default alerted to the possibility of making a *who* decision when posting something. This is done via a choice box placed right next to the post-input field (see Figure 12.2). When opened, it provides information about possible recipients and also arranges these options hierarchically, so that the number of choices remains manageable.[7]

In view of the challenging nature of this choice task, which is a type of configuration problem, it is understandable that tools have been developed which aim to make it tractable by applying several of the ARCADE strategies in conjunction. We will use as an example the tool FREEBU (Gao et al., 2012[8]), which helps people create lists of their FACEBOOK contacts.

Figure 12.3 shows one of FREEBU's several modes that largely supports attribute-based choice: Contacts are grouped on the screen in terms of attributes such as home country, university attended, and whether they know a particular language. The user's choice task is not, as would be typical of attribute-based choice, to select one or more individuals who are in some respect desirable but rather to compose a new list of friends that is likely to be useful in future choice situations (e.g., when the user is deciding whom to share a particular media item with).

This *column mode* applies ARCADE strategies as follows:

- *Represent the Choice Situation:* The most salient contribution of the interface is the way in which it organizes the relevant items (here: persons) on the screen in a way that facilitates a particular way of generating options (here: possible lists of contacts). Given that members of a useful contact list are often similar with respect to particular attributes, organizing the contacts by at-

[7]The top-level option "custom", which appears to be rarely used, allows for arbitrarily complex combinations of the basic elements.

[8]`http://people.cs.kuleuven.be/~bo.gao/freebu`

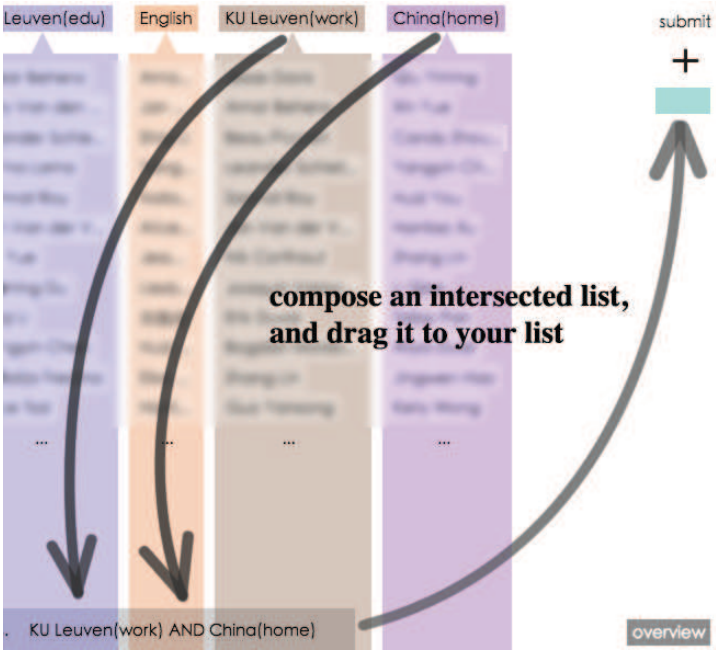


Figure 12.3: An image from the FREEBU interface for organizing FACEBOOK friends into lists. (This image, which obfuscates the friends' names for privacy reasons, offers users advice about how to use the interface to create a list.)

tribute makes it relatively easy for the user to see promising ways of grouping contacts.

- *Access Information and Experience:* Though it is not the primary purpose of this visualization to convey new information to the user, it can happen that the user becomes aware of relevant facts about his friends that he had not previously noticed (e.g., the fact that one or more of them comes from a particular town).

- *Combine and Compute:* Creating a particular arrangement of items on the screen requires a certain amount of computation—which is quite straightforward in the column mode but more sophisticated in other modes of FREEBU, such as the one that clusters friends on the basis of their relationships to each other.

- *Advise About Processing:* The user interface includes some textual hints as to how to go about creating a friend list. In principle, these hints could be formulated without reference to specific user interface elements (e.g., "Look for sets of friends that resemble each other with respect to two or more attributes ..."). In the FREEBU interface, as in many other such interfaces, procedural advice of this sort is blended with instructions and hints about how to operate the user interface: "You can also first create an intersected column by dragging multiple columns into the 'intersection' box at the bottom of the screen; only the friends who share the common attributes of these columns will stay in the intersection box. Then drag the intersection box into the target list."

12.4 Socially Based Choices About Privacy

Given that privacy-related actions in essence involve more than one person, it is natural that social expectations and examples play an especially important role (see, e.g., Greene et al. [2006]). The importance of social norms is underscored by the often striking differences that exist among cultures with regard to particular types of disclosure (see, e.g., Petronio, 2002, pp. 40–42). Palen and Dourish [2003] note that "privacy management ... involves combinations of social and technical arrangements that reflect, reproduce and engender social expectations, guide the interpretability of action, and evolve as both technologies and social practices change" (p. 133).

These examples and expectations may concern either the *disclosure* of personal content (e.g., the revelation of intimate personal details in a blog post) or the *accessing* of such content that belongs to others (e.g., "shoulder surfing" to look at the email inbox of a colleague that is displayed on her screen).

In the social networking site scenario, privacy-related social expectations can be expressed explicitly, as through a community policy statement, or implicitly through the behavior of individual members (cf. 11.4). An example of the former case is provided by the "Com-

munity Guidelines" page of the popular diet and fitness community MYFITNESSPAL,[9] which includes, for example, precise rules (i.e., articulated social expectations) concerning what may and may not be shown in photos that are posted by members who want to show off their progress in improving their physique.

As was discussed in connection with online communities (11.4), when the only available evidence is community members' behavior, it can be hard for a member to know whether what they are dealing with is an expectation or merely examples that are not necessarily considered appropriate. The behavior of other persons that a chooser interprets can consist of (a) examples that are not causally related to his own behavior or (b) responses to his own choices (in which case the chooser can be seen as applying the trial-and-error-based pattern to acquire knowledge about social examples and expectations).

Both cases are illustrated in a lengthy message board thread on MYFITNESSPAL[10] in which members have contributed more or less revealing and embarrassing sequences of "before and after" pictures of themselves to illustrate their progress in losing weight. These contributions often evoke evaluative comments (e.g., "I love these! It's amazing how people change"). The pictures and the comments that they evoke can be seen to influence the subsequent contributions of other members (e.g., "Wow these are awesome! Great job everyone! I feel ready to do a new fat face comparison pic!"). Note that it's not clear in these examples to what extent the evaluative comments refer to (a) the weight-losing progress shown in the photos or (b) the contributor's willingness to display these photos. This fact illustrates again the point made in 12.1.3 about how privacy-related choices are often "bundled" with other aspects of a person's behavior.

12.5 Policy-Based Choices About Privacy

Although the term *policy* is relatively seldom applied in the privacy area in the sense of the ASPECT model's policy-based choice pattern,

[9] `http://www.myfitnesspal.com/welcome/guidelines`

[10] Specific reference omitted here for privacy-related reasons.

it is of course possible for a user to have a privacy-related policy in the sense of ASPECT: one or more rules or principles that specify (directly or indirectly) what privacy-related choice will be made in each of a set of situations, possibly as a function of particular parameters of the situation. For example, a user could have one of the following rules: (a) She will never give her birthday to any website. (b) She will never give her birthday when it is possible that it might be used for direct marketing. (c) She will give her birthday to a social networking site but will allow only her friends to access this information.

To be sure, arriving at a set of meaningful policies can be a challenge for the individual user in such a complex domain. This difficulty is presumably one reason for the abundance of privacy-related advice that can be found on the internet—such as the brief article titled *7 Things to Stop Doing Now on Facebook* (Consumer Reports Magazine, June 2010), which in effect suggests seven rules that a FACEBOOK user might adopt as part of his privacy-related policies. Personal policies can also be derived from social expectations of the sort discussed in 12.4.

One advantage of policy-based choice in this domain is that it makes it unnecessary for the user to think continually about all of the various privacy-related problems that might arise—which, as we have seen, can be hard (and possibly unpleasant) to anticipate and evaluate. Instead, she can focus on sticking to her policies, assuming that if she does so everything will probably be all right.

Another advantage of policies is that it is sometimes possible to delegate their execution to some sort of automated agent, which involves applying the strategy *Evaluate on Behalf of the Chooser* (cf. Section 9). This prospect is especially attractive in the privacy domain, since in principle privacy-related choices can arise at the rate of one every few seconds, as when a user is surfing the web or moving around the city with a location-aware app. It is therefore understandable that many interactive systems provide forms in which a user can in effect specify policies for making choices such as those about what personal data should be exposed in particular situations.

An example of such a form in a well-known privacy protection agent is shown in Figure 12.4. This web-based form (to be discussed in more

detail below) differs from most forms of this sort in that the user does not specify what the system is supposed to do autonomously but rather when the system is supposed bring the user into the loop by warning him of the danger of a privacy violation.

Before discussing research on this type of policy specification, we should digress a bit to analyze a frequently used concept that tends to obscure understanding in this context: that of *privacy preferences*.

12.5.1 Note on the Concept of "Privacy Preferences"

As can be seen in Figure 12.4, forms that allow a user to specify a policy of this sort are often labeled with the word "privacy preferences". This term, which is widely used in the HCI privacy literature in other contexts as well, is more misleading than helpful, because the term *preferences* can be interpreted in at least three ways:

1. As the specific choices that a person makes in particular situations (e.g., "I can make my medical information available to this website now, but I prefer not to").

2. As a chooser's more *general predispositions* that influence her choices (as in "My general preference is not to make my medical information available to websites").

3. As the choices that a person makes about *how to fill in a form* of this type (as in "I could have checked the box asking whether my medical information should be made available to websites, but I preferred not to").

When the term "privacy preferences" is used in interactive systems or scientific discussions, it is rarely made clear which of these meanings is intended. A consequence is that the general impression is created that these three concepts are tightly correlated:

- A. that a user's specific privacy-related choices are (largely) determined by general predispositions.

- B. that these predispositions are accurately captured by forms that ask about "privacy preferences".

With the background of the ASPECT model, we can see that neither of these implicit assumptions is even approximately accurate:

Regarding assumption A: The choices that people make are not always determined primarily by general predispositions of these sorts. For example, within the experience-based pattern the chooser may simply repeat a previous choice that the current situation reminds her of. Within the socially based pattern, she may imitate a momentarily salient social example. When applying the consequence-based pattern, she may anticipate specific consequences that can arise because of unique aspects of the current situation; and her evaluation of these consequences can depend on temporary factors such as the relative momentary salience of her various goals.[11]

Regarding assumption B: Where more general predispositions are involved, they can take qualitatively different forms, some of which are not naturally captured by preference forms. Here are some examples in terms of three of the ASPECT choice patterns:

- Within the policy-based pattern: a personal policy that applies to some class of choices (e.g., "I will never associate my photo with a social network profile").

- Within the experience-based pattern: a habit of declining any opportunity to upload a photo to a profile; or a strong negative affective association with this action, which may have been acquired through bad previous experience.

- Within the consequence-based pattern: A general belief that performing this action can lead to serious negative consequences.

[11] In the relevant psychology literature, the term *construction of preference* is sometimes used to refer to phenomena like these. See Lichtenstein and Slovic [2006] for an influential collection of articles that are relevant to the question discussed in this section. But even in this literature, terms like *preference* and *construction* are rarely explicitly and clearly defined. See also the book by Hausman [2012] for a thorough discussion of the concept of *preferences* from the perspectives of economics, psychology, philosophy, and everyday language. The very existence of a book like this should discourage anyone who employs this term from assuming that readers will have a clear and accurate idea of what he is talking about.

By suggesting a tight correlation among points 1–3 above, the use of the term *privacy preferences* serves as a sort of smoke screen in discussion of privacy-related choices, preventing us from seeing clearly how these choices come about and what we can do to support them. Researchers and practitioners in this area would do well to adopt the policy of never using this term—even though applying this policy may at first require a great deal of self-control and investment of the effort required to figure out what is really being referred to. It might help to remember the following slogan:

> People don't "have preferences"; they *make choices*.

12.5.2 Studies of Policy Specification

Returning to the more specific topic of "privacy preferences" forms, we can see that, when a user fills in such a form, she cannot in general be assumed to be reporting straightforwardly on a policy that already exists in her mind. Rather, she is being asked (or required) to choose, from among a set of possible policies that the system can enforce on her behalf, the policy that seems best suited for automatically making choices that she would find appropriate. This task is similar to the subtasks of formulating and evaluating a policy that were discussed in 9.4 and 9.5, the main difference being that in this case a set of possible policies is being provided from which the user has to choose.[12]

Against this background, the question of how natural and useful a given "preferences form" is found by a given user is always an empirical question. A number of studies have looked at questions of this type.

One of the best-known and most realistically tested systems for automatically applying users' privacy policies is the Privacy Bird of Cranor and colleagues (see, e.g., Cranor et al., 2006), which was already mentioned in 12.2.1. This system took advantage of the fact that many websites made their own "privacy policies" concerning their use of personal data available in a formal XML-based language defined by the Platform for Privacy Preferences (P3P).[13] The Privacy

[12] For more general discussions of the question of the subtleties involved in eliciting people's "preferences" and values, see Fischhoff [2006] and Fischhoff [1991].

[13] `http://www.w3.org/P3P`

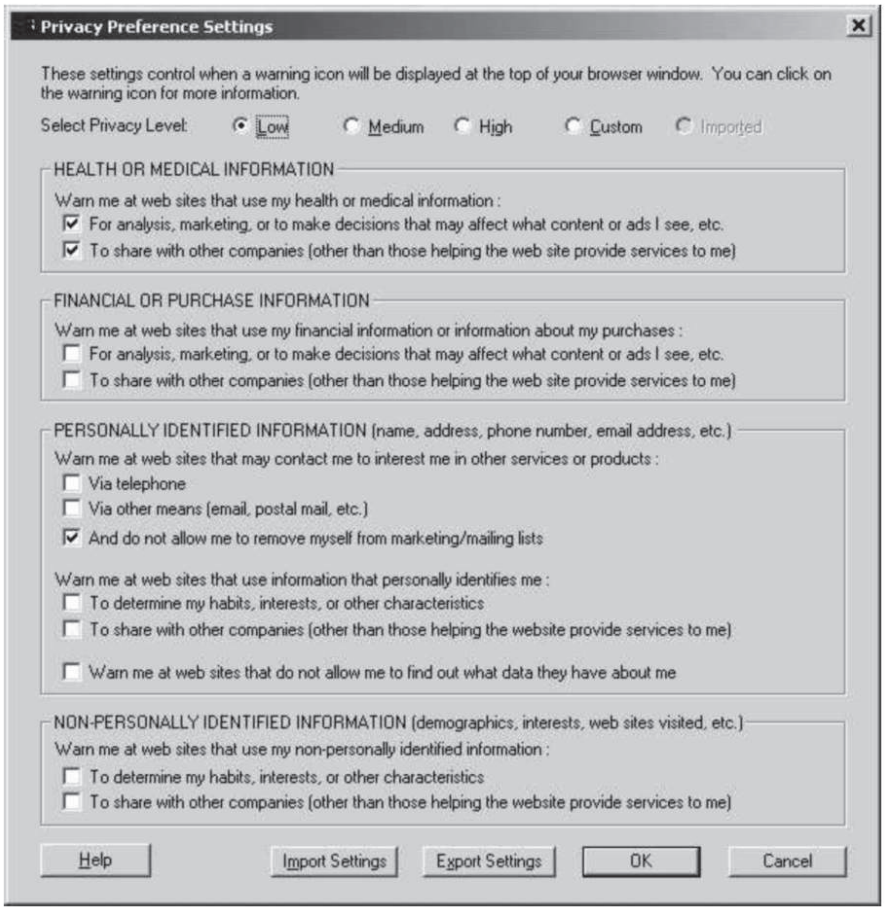


Figure 12.4: The screen used by the PRIVACY BIRD to elicit a user's policy with regard to privacy-related choices. (From `http://www.privacybird.org/tour/1_3_beta/tour.html`.)

BIRD created a similarly formal representation of the privacy policy that a user had specified via the form shown in Figure 12.4. The PRIVACY BIRD could then automatically determine, whenever the user visited a website with a formalized privacy policy, whether the user

should be warned (by a red bird) about possibly undesirable disclosure of personal data.[14]

P3P is an interesting case for the current discussion because, despite its elegant technical design and high hopes and extensive support in its early stages, it was not widely adopted by websites and users and was ultimately suspended by the relevant W3C working group. A variety of explanations have been offered, some of which serve to remind us of the challenging context that confronts attempts to support privacy-related choices (e.g., insufficient support by browser implementers, the ease of circumventing the Privacy Bird with invalid site policies, the lack of enforcement through legal or self-regulation, and user interface problems).[15] The problem most directly relevant to choice architecture is the difficulty that users have in specifying their personal privacy policies. As Cranor et al. [2006, pp. 7–9] explain, the user may want information disclosure decisions to depend on considerations that cannot be expressed in the policy specification form that is provided (e.g., the branch of industry to which the website belongs or the specific companies with which information will be shared). They are also faced with the more general difficulties of policy formulation and evaluation that were discussed in 9.4 and 9.5, which are especially acute in a domain in which the consequences of actions can be so hard to anticipate.

Lederer et al. [2004] describe some of these difficulties in the privacy context as follows:

> The act of configuring preferences is too easily desituated from the contexts in which those preferences apply. Users are challenged to predict their needs under hypothetical circumstances, and they can forget their preferences over time. If they predict wrongly, or remember incorrectly, their configured preferences will differ from their *in situ* needs, creating the conditions for an invasion of privacy. (p. 447)

[14] Other P3P-based user agents performed privacy-related actions automatically, such as deciding whether to block cookies or whether to allow access to a user's electronic wallet (cf. Cranor et al., 2006).

[15] For a recent brief survey and further references, see Morton et al. [2013].

In a similar vein, Berendt et al. [2005] suggest a number of reasons why people's privacy-related behavior in e-commerce sites can deviate from the policies that they express in questionnaires.

Although the results and considerations just summarized highlight a number of problems with the automation of privacy-related policies, there will presumably always be some efforts to achieve this type of automation at least in connection with choices that arise with high frequency and seem to lend themselves to policy-based choice. One question that arises in this context is whether a policy specification interface should allow the specification of complex rules, referring to a number of variables; or whether only simple rules are needed (cf. Iachello and Hong, 2007, pp. 47–49). As one might expect, there appears to be no single answer that applies to everyone. Using an experience sampling methodology in a field study, Anthony et al. [2007] asked participants on a number of occasions whom they would be willing to share their current location with. The responses of most users could be captured with the simple rules "Never share your location with anyone" or "Always share your location with the people on your white list". For a handful of participants, however, the choices depended on characteristics of the situation in which they found themselves, which implies that they would require a relatively expressive language to formulate a satisfactory policy for revealing their location.

Benisch et al. [2010] conducted a somewhat similar study in which they noted that many users would require considerably more complex policy specification interfaces than the most commonly available type (i.e., a whitelist of individuals with whom the user is always willing to share) if they wanted to be sure that the system would make roughly the same choices that they themselves would make on a case-by-case basis. The authors also took into account the fact that specifying a complex policy can require considerable effort and that at least some users might therefore choose to adopt a relatively simple policy even if it didn't match their case-by-case choices.

In view of the effort and difficulty associated with manually specifying policies of this sort, it is natural to consider computational support for this process. Fang and LeFevre [2010] presented and evaluated a

Privacy Wizard that uses sophisticated machine learning techniques to construct a policy on the basis of a limited amount of input from a user as to whether he would reveal a particular piece of information to particular individuals or groups of people. They show how the recommended policy can be visualized to the user so that he can better understand and evaluate it. This approach involves applying the strategies *Combine and Compute* and *Evaluate on Behalf of the Chooser* to the subtask of formulating a policy. It will be interesting to see, in future research, to what extent and in what contexts machine learning can help users to arrive at privacy-related policies whose automatic application will yield satisfactory results.

12.6 Trial-and-Error-Based Choices About Privacy

Especially because of the inherent difficulty that users have in anticipating the likely privacy-related consequences of their actions, it is natural—though, as we will see, problematic—for privacy-related choices to be based on trial and error. As was discussed more generally in 10.4, the information and experience that the chooser acquires through trying options out can take various forms.

12.6.1 The Difficulty of Learning From Privacy Violations

Let's consider first the most obvious form: feedback about the significant consequences of the chooser's actions. Among the most significant consequences are violations of the chooser's privacy. Unfortunately, there are several reasons why learning from this type of consequence tends to be problematic.

First, personal data is an information good; once it is out, it cannot be taken back. There can in general be no "undo" button for any specific piece of information (cf. 10.1.2). Even though a "right to be forgotten" has now found its way into the draft of the new European Union privacy directive, this right is associated with many difficulties, including technical ones (see, e.g., Druschel et al. [2012]). The irreversibility of possible consequences is an important property with regard to exploration strategies, especially when the consequences in question are seri-

ous. This fact is taken into account in some user interfaces (e.g., when data are about to be permanently deleted) with warnings like "This action cannot be undone!", which tell the user, among other things, that simply trying the action out is unlikely to be part of a good exploration strategy. These warnings can be viewed as an instantiation of the strategy *Access Information and Experience* insofar as they inform the user about a property of the consequences of an option; if they are presented emphatically as warnings, they can also serve as advice to think carefully before choosing the option (*Advise About Processing*). These tactics may deserve increased attention where privacy-related choices are concerned. The privacy nudges of Wang et al. [2013] mentioned in can serve, in somewhat different ways, to discourage casual trial and error.

Even when viewed purely as sources of information, privacy violations tend to be relatively hard to process as informative feedback. As was indicated in 12.2, the *prediction* of the results of privacy-related choices tends to be made more difficult by the fact that consequences are often interrelated, invisible, indirect, delayed, and nondeterministic, partly because of the involvement of multiple stakeholders. These same factors also constitute obstacles for the trial-and-error-based pattern by making it hard for a chooser to recognize after the fact what *has* happened as a result of a privacy-related choice. For each of the pitfalls in the interpretation of feedback that were discussed in 10.5, the interested reader should be able to think of examples in the privacy domain. For instance, many privacy violations are undetectable for most users, such as those in the surveillance scenario that involve the reading and analysis of data from servers or transatlantic cables.

12.6.2 Learning From Information About Potentially Problematic Situations

Because of all of these problems with learning from privacy violations, a promising approach is to provide to a user feedback about *potential* privacy threats that have arisen because of her actions but which have (mostly, at least) not resulted in actual problems (cf. the more general discussion in 10.4 of the approach of providing feedback about states

that are likely to be correlated with important outcomes). Specifically, it can be helpful for a user to know who has been able to acquire what information as a result of his actions—and what they may have been able to infer from this information. Even if all of the observation and inference that the user becomes aware of does not lead to any privacy violation, this information can give an indication of the danger of such violations if the user continues to make the same choices.

Providing this type of feedback takes different forms depending on whether the other "knowers" are ordinary individual users (e.g., social network peers) or institutions.

Feedback About Observation by Individuals

System and interface design can help a user to find out about the identities and activities of other persons who have become knowers of that user's private information. An increasing number of "How To" web pages, apps, and plugins enable users to "observe the observers".

An example is an app for seeing who has viewed your FACEBOOK profile.[16]

Several data-based privacy awareness tools (see the general comments in 12.6.1) summarize, visualize, and score a user's past publishing behavior. An example is the app PRIVACYCHECK (Figure 12.5), which shows what types of personal information from the user's FACEBOOK pages are made available by the FACEBOOK API to websites that the user visits.[17]

An example of a system that was designed to help users learn by trial and error when to disclose their locations while on the move is the LOCYOUTION system of Tsai et al. [2009]. The user can specify, for each weekday, the times of day during which a particular class of contacts (friends, acquaintances, or strangers) are allowed to see her location. She can also check, via a special feedback page, which persons have tried to see her location at what times and whether that person was allowed to do so by the rules that were defined at the time. In this way, the user can essentially debug the rules over time—though the

[16] `https://www.facebook.com/WhoHasSeenYouProfileNewApplication`
[17] `http://www.rabidgremlin.com/fbprivacy/`

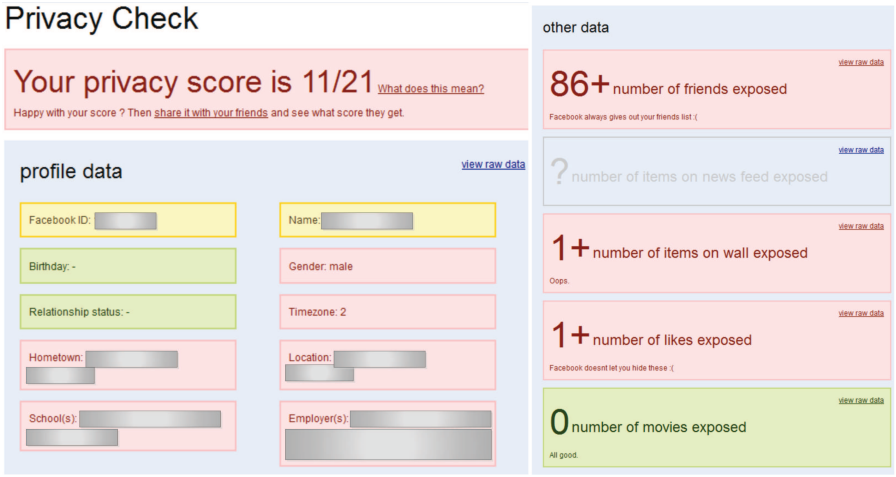


Figure 12.5: Screenshots from the FACEBOOK app PRIVACYCHECK. (The values of some fields are masked in this figure for privacy reasons.)

value of the resulting policy can be limited by the expressiveness of the mechanism for specifying rules (cf. the discussion in 12.5.2). Although the participants in this study by Tsai et al. [2009] expressed general satisfaction with the rule specification mechanism, they also suggested several additional types of rules that they would have liked to have.

A possible drawback of providing even excellent feedback about the responses of one group of persons (e.g., peers such as other users of a social-networking site), is that it may divert the user's attention away from all those others who are "watching" (cf., the discussion, in earlier subsections, of social privacy vs. institutional privacy and surveillance). This phenomenon can be seen as a case of (maybe inadvertently) biased representation of the choice situation—specifically, the provision of a biased sample of information about who is observing the chooser's actions.

Feedback About Tracking by Organizations

One way to reduce the bias just mentioned is to make available to ordinary users tools for accessing information about what organizations are

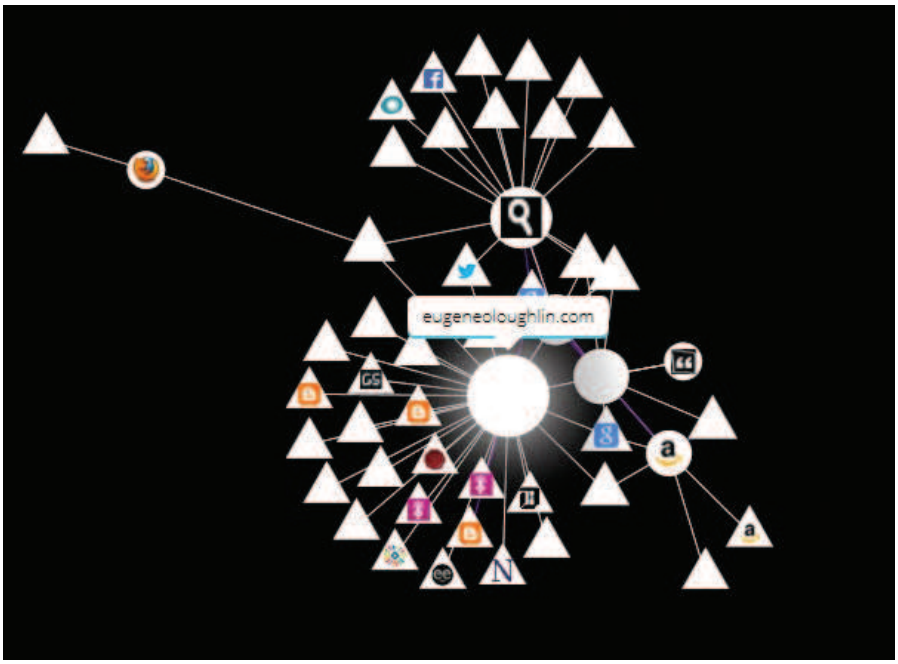


Figure 12.6: Screenshot from the LIGHTBEAM plugin for Mozilla FIREFOX. (From eugeneoloughlin.com.)

monitoring their behavior—and what knowledge they can derive from this monitoring. An example is the plugin LIGHTBEAM[18] (Figure 12.6) for the FIREFOX browser, which enables a user to get an overview of the websites that track his website visits—including cases where one website has recorded his visits to other websites. Typically, users are surprised to see how many sites are tracking their behavior and how much information a single site can acquire.

12.6.3 Characterization in Terms of the ARCADE Model

The feedback provision techniques just discussed are representatives of a large class of choice support techniques that instantiate a combination of the first three ARCADE strategies:

[18] `https://addons.mozilla.org/en-US/firefox/addon/lightbeam`

- *Access Information and Experience*, in that information about consequences is being provided;
- *Represent the Choice Situation*, in that this information needs to be represented in a convenient way for the chooser to be able to make use of it; and
- *Combine and Compute*, in that in general some computational processing of the relevant information is required to generate a suitable representation.

12.7 Concluding Remarks on Privacy-Related Choices

In this section, we have given a number of examples of existing or possible tactics for supporting privacy-related choice. But a more important goal of this section has been to shed light on the special challenges that are raised by privacy-related choices, by discussing them in terms of the choice patterns of the ASPECT model. Though we hope that those who design and deploy relevant computing technology will benefit from being able to think about these challenges in terms of the ASPECT choice patterns and the ARCADE strategies, we believe that the challenges are too great to be dealt with completely even by the best-informed interaction and system design. Equally important is the goal of conveying to users of computing technology more realistic mental models of privacy-related choices (including their effects on other people) and a grasp of the options and tools that they have available for dealing with them (cf. 4.7.2).

7

PAPER 5:

Learning analytics and their application in technology-enhanced professional learning

CHAPTER 13: LEARNING ANALYTICS AND THEIR APPLICATION IN TECHNOLOGY-ENHANCED PROFESSIONAL LEARNING

Bettina Berendt, KU Leuven, Belgium

Riina Vuorikari, **Error! Bookmark not defined.**European Schoolnet, Brussels, Belgium

Allison Littlejohn, Glasgow Caledonian University, UK

Anoush Margaryan, Glasgow Caledonian University, UK

Learning analytics (LA) is the "measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs" (Siemens & Gasevic, 2012). Originally, “analytic” refers to a way of using data to support decision-making and understanding a domain. Essential LA components are (1) data, (2) goals or (research) questions, optionally based on educational theory, (3) measures that give information about goal attainment or (research) construct, optionally (4) descriptive or predictive models that use these values as variables, and (5) computing models and routines that compute these measures’ values, modelling results from the given data. LA systems also comprise (6) automatic or semi-automatic ways of reporting these results to the chosen stakeholders. Optionally, (7) the results can be deployed within some application functionality. Examples of 1) and 3) are ‘clickstream’ data[i] used to measure learner behaviour and knowledge, or text data underpinning domain models. The goal (2) could be to depict collaboration between learners. The descriptive or predictive models (4) may comprise learner profiles or models for predicting whether a learner is ‘at risk of dropping out’. The computing models that determine these measures (5) range from simple counts via clustering techniques to classifier learning.[ii] The models may be purely statistical (correlating measured variables) or refer to theory (which would, for example, explain why someone with certain behaviour is at risk of dropping out, and what the behaviour and the

risk have to do with learning). A typical choice for (6) is dashboards, when the results are reported to teachers or information is given as feedback to learners. An example of (7) is the use of learner models to offer users personalised learning resources that are assumed useful for an individual's learning. For a related model of components, see (Greller & Drachsler, 2012).

Learning analytics is developed within various disciplinary communities and is an instantiation of different ideas, each with its own long tradition. "Business analytics on e-learning" is one such perspective; another one is "Web analytics on e-learning". Other traditions are those of "learning analytics", "knowledge analytics" and "academic analytics" developed in the field of Knowledge Management (see Siadaty et al., 2012 for a use of traditional Knowledge Management concepts in LA for workplace learning). A common theme in these fields is the detailed analysis of behavioural data describing the usage and production of knowledge resources. LA is also closely related to *educational data mining* (EDM). Both analyze learning data: EDM focuses on the data-mining models and fully-automated modelling and personalisation, whereas LA draws on models from different disciplines. LA also applies modelling as well as personalizing in a more semi-automated and interactive fashion (Siemens & Baker, 2012). Lastly, interactive LA, that is designed to be used by learners and others who are directly involved in the learning, are instances of *feedback and awareness systems* (Berendt et al., 2012): systems that give a user feedback about her/his own behaviour, in an attempt to raise awareness about issues such as how one learns or how one learns in relation to what one thinks about how one learns.

Learning analytics could be valuable in workplaces that focus on effectiveness and operational excellence (van Barneveld, Arnold, & Campbell, 2012). In this chapter, we

investigate this claim in more detail, highlighting both the potential and issues of applying LA in technology-enhanced professional learning. In the remainder of this section, we offer a brief overview of important themes in LA research and practice. Then, in Section 2, a case study illustrates the use of LA within a pan-European platform for teachers' professional development. Section 3 highlights challenges, and Section 4 concludes with an outlook on key issues in the use of LA in supporting professional learning.

The field of LA is motivated to a large degree by the recent growth in the use of Learning Management Systems and other online environments, and the wealth of data they produce. In such environments, learners use digital tools and leave digital traces. There is a sense that these traces can be used to make learning more effective, yet it is not always clear how the data can best be used. Additionally, there is the growing pressure on academic education institutions for accountability through performance measurement (Ferguson, 2012). In this sense, the unique features that LA strive for is to be learner-centric and informed by pedagogical theory (Ferguson, 2012). LA are also often interactive and visual in their reporting of results, in particular when learners are the recipients of the data, cf. (Bienkowski, Feng, & Means, 2012; Santos, Govaerts, Verbert, & Duval, 2012). The focus on 'interactivity' stems from the goal of supporting reflection (McAuley, O'Connor, & Lewis, 2012), metacognition, and thereby self-regulation, as key drivers of professional learning (Littlejohn, Milligan, & Margaryan, 2012; Siadaty Jovanović, & Gašević, this volume). Analytics may comprise data on learning activities, such as the number of posts made in a forum, comparisons of these to average levels in a relevant community, and interpretations of these activities in terms of goals or categories, for example, reasoning, evaluation, extension and challenges (Ferguson & Buckingham Shum, 2012).

In today's networked environments, learning rarely happens in isolation (Sloep, this volume). *Social Learning Analytics (SLA)* draws on the idea "that new skills and ideas are not solely individual achievements, but are developed, carried forward, and passed on through interaction and collaboration […] Understanding learning in these settings requires us to pay attention to group processes of knowledge construction – how groups of people learn together using tools in different settings. The focus must be not only on learners, but also on their tools and contexts" (Ferguson & Buckingham Shum, 2012). When SLA are used, the unit of analysis - and the potential consequences - differ. For example, a learning group with members who are at risk of dropping out should receive help *as a group* (rather than only support for the individual members). This solution can encourage a mutual awareness of problems and a shared approach to support that overrides opinions about individuals' failure or the abandonment of responsibility towards others.

In formal learning settings, such as professional training or academic education, learning analytics most often focus at the level of a 'course' or another administrative structure (Ferguson, 2012). Whereas in informal learning settings, LA focus at the level of groups and networks of learning (ibid). In informal learning settings, where there is usually no set syllabus, course structure, and accreditation, learning interactions are not usually guided by teacher-learner relationships. One example of an informal, professional learning setting is a 'learning network'. Learning networks are technology-supported communities, in which learners share and develop knowledge (Sloep & Berlanga, 2011). Studies of the benefits of such networks for professional development (OECD, 2009; Suthers, Dwyer, Medina, & Vatrapu, 2010; Berlanga & Vuorikari, 2012) highlighted a need for tools and methods that could help find "reliable evidence of how, when and why online social networks do, and do not, advance learning" (Schlager, Farooq, Fusco, Schank, & Dwyer, 2009). Examples of LA

methods and processes that could be applied in informal learning setting have been proposed (Song, Petrushyna, Cao, & Klamma, 2011; Cambridge & Perez-Lopez, 2012; Vuorikari & Scimeca, 2013).

We present one such example, from a technology-supported teacher network called eTwinning[iii]. With more than 190,000 members, eTwinning is a European network of schools. Participating teachers collaborate online while learning new skills. We first introduce the context for the case study – the need for teachers' professional development, before we present the concept of eTwinning Analytics to exemplify the use of SLA in a professional context.

Case study: eTwinning analytics

Nowadays, there is a need to upskill K-12 teachers to help them respond to rapidly changing needs in society. Teachers, however, feel that they do not have sufficient opportunities for professional development (OECD, 2009). Today, the most common form of ICT-related professional development undertaken by teachers is "personal learning on ICT in their own time" (Wastiau et al., 2013). Only one out of three students in Europe are taught by teachers who have participated in compulsory ICT training (Wastiau et al., 2013). Co-operation amongst teachers can create opportunities for exchange of ideas and practical advice, enhancing professionalism, increasing feelings of self-efficacy and preventing stress and "burnout" (OECD, 2009).

Teacher networks have started to emerge, with early examples including Tapped-in[iv] and Teachernet[v]. These aim to improve both the *quality of the teaching profession* and the

learning experience of students, by encouraging collaboration and knowledge exchange at both teacher and student level (Vuorikari et al., 2012). Such networks allow teachers to upskill and to gain new competences in the context of daily work. However these networks rely on professionals' voluntary participation.

eTwinning is one such network. The eTwinning platform offers teachers three main streams of activities: (1) school collaboration projects, where teachers can find partner schools to run cross-border activities using ICT; (2) various formal and informal professional development (PD) opportunities, including online courses and special interest groups; and (3) social networking. eTwinning needed longitudinal studies to monitor and measure various forms of teacher co-operation, a need that led to the development of eTwinning analytics. eTwinning analytics is "the measurement, collection, analysis and reporting of data about eTwinners and their contexts, for the purposes of understanding and optimising their co-operation and the environment in which it occurs*"* (Vuorikari & Scimeca, 2013). eTwinning Analytics fall into the category of Social Analytics, where the interest is in teachers' co-operation behaviours and patterns over a long-term period (5 years) and how these patterns and behaviours can support teachers' continuous professional development, when knowledge building takes place in a cultural, social and technology-enhanced setting.

The components of the eTwinning Analytics are as follows: (1) Data are gathered from the eTwinning platform. (2) The goal is to operationalise the construct of teachers' co-operation in eTwinning - or in other words, to find quantitative measures of what it means for teachers to cooperate and what it means when one individual cooperates more than another (3). Measures include teachers' actions using various digital tools, as well as their cooperative interactions with each other. These cooperation activities are mapped using OECD indices,

namely that of a) teachers' *exchange and co-ordination activities*, for example the exchange of learning materials and ideas, and b) *professional collaboration activities*, such as cross-border school collaboration. In the following example, we demonstrate the use of eTwinning Analytics to explore three research questions (RQ). An extended version of this study can be found in (Vuorikari & Scimeca, 2013).

RQ1: eTwinning retention rate: Is there evidence of teachers remaining engaged with eTwinning over a long period of time (i.e. since its start in 2005)?

"Retention rate" is a Web-analytics measure used in online marketing. It is the percentage of users who sign up for the service and come back within a period of time. The retention rate for eTwinning refers to the percentage of teachers who have registered on the platform since its inception in 2005, and who still return to log-in annually. Figure 13.1 shows the eTwinning retention in 2011 and in 2012. The x-axis represents the number of years since registration on eTwinning, and the vertical axis represents the percentage of teachers. "0 years" refers to people who registered in 2011, "1 years" to people who registered in 2010, etc.

[PLACE FIGURE 13.1 HERE]

We can observe that in year 0, the retention rate is high. For example, 89% of users who registered on eTwinning in 2011 and 86% who registered in 2012 returned to login into the platform at least once during that year. A year after the registration, we can observe a steep decline: about 40% of users still login onto the eTwinning. This trend remains much the same from 2011 to 2012. Finally, about 1 in 6 of teachers registered 5 to 7 years ago still remain engaged. Therefore, it seems that eTwinning has the potential to engage users over a long

time-period. More research is needed to understand why so many teachers drop out in Year 1 only after a short involvement. What challenges did they face? Were these related to the collaborative nature of their activities on the platform? Is the drop-out related to problems with the platform, or did the network not help teachers sufficiently with their professional learning needs?

RQ2: Teachers' co-operation activities: Are there any trends that emerge in teachers' co-operation activities over a long period of time?

Figure 13.2 shows the extent to which teachers have engaged in various co-operation activities, including participation in cross-border school collaboration and social networking activities, such as adding Contacts and/or participating in Teachers' rooms. 'Contacts' are explicit links to other users as in social-networking sites, and 'Teachers' rooms' are interactive spaces dedicated to various subjects such as "Les langues romanes", a French-speaking room on Romanic languages. The number of years since registration are counted backwards from 2011 as above in Figure 13.1, and the percentages illustrate a data snapshot taken in February 2012.

[PLACE FIGURE 13.2 HERE]

Two patterns can be observed. Firstly, in terms of collaboration, it appears that in the early years of registration on the platform users are less engaged (average 18%) in joint project work compared to those who have been on the platform for more than 2 years (average 30%). Secondly, in terms of social networking, teachers in their early years of participation in the platform (registered in year 0 and 1) are slightly more involved in the Teachers' rooms than those who have been using the platform for longer. Similarly, the Contacts feature is used by

almost half of the teachers in their first year of registration (45%), and use seems to intensify after that starting period.

Observing longitudinal patterns is important in helping us understand how a digital platform such as eTwinning can serve teachers' professional learning needs over the length of their career. To experience a full range of professional development activities in eTwinning, and to gain full advantage of the participation in a teacher network, teachers need to make a substantial time investment. The analytics show that teachers in their early years of participation in the platform are less engaged in co-operation activities, a finding that prompts further questions of how the platform can support them better. The monitoring of teachers' professional development paths through eTwinning Analytics is outlined in Cao, Klamma, Pham, & Vuorikari (2012).

RQ3: Use of social networking tools: Do teachers who engage in professional collaboration on the platform and those who do not use social networking tools in the same way?

[PLACE FIGURE 13.3 HERE]

Figure 13.3 illustrates the usage of four different social networking tools (Contact, Profile picture, Journal Wall posts and Teachers' rooms) by those who engage in cross-school project work and those who don't. The Contacts and Teachers' rooms functionalities have been described above. Profile pictures are pictures, usually photographs, uploaded by the users to describe themselves. Journal Wall posts are short descriptions of current activities, also uploaded by the user. The percentages shown describe the same data snapshot taken in February 2012 as used for Figure 13. 2.

Social networking tools appear to be used more by teachers who are involved in project collaboration (average 64%) than those who are not (36%). An exception is the Contacts tool, which is used by a similar proportion of those who do and those who do not engage in the project, collaborating within the platform. However, the Analytics, as such, do not shed any light on why this is the case.

About 2 out of 3 users of social networking tools on the eTwinning platform were also active in project collaboration. The results illustrate that teachers use a large variety of tools and engage in many activities through the platform. However, eTwinning Analytics cannot give insight into the cause and effects of tools usage and professional collaboration. Questions remain about professionals' interactions with various tools. We are interested in understanding, for example, whether the use of social networking tools can lead to better project collaboration opportunities. However, one limitation is the difficulty in measuring indicators of communication taking place outside the platform. These limitations pose challenges for gaining a good, overall picture of the interactions and learning contexts within tools such as the eTwinning network. In the conclusions section in this chapter, we will sketch some possible ways to overcome these limitations.

The teachers who are not involved in project collaboration via the platform are nevertheless building weak ties through social networking (such as, sourcing and adding other teachers as professional contacts). Weak ties play an important role in the enhancement of information flow in networks, leading to emergence of new ideas (Haythornthwaite, 2001). Previous studies on eTwinning networks have evidenced that both Project and Contact networks are dense and well-connected, illustrated by the number of edges, average path length, diameter,

the number of components, and other measures of connectedness properties in these networks (Pham, Cao, Petrushyna, & Klamma, 2012).

Challenges

In the previous section, we demonstrated a 'proof of concept' of SLA to support learning within an informal, professional network for teachers. Instead of focusing on each individual learner, the goal was to use aggregate statistics to view emerging, long-term trends. In general, the use of descriptive statistics does not allow us to distinguish between cause and effect. However, the results allow the development of new hypotheses for further investigation. More sophisticated and combinatorial analysis of data will lead to a deeper understanding of when informal learning networks better support teachers' personal and professional development goals. For example, mixed method analysis (using exploratory and confirmatory methods and/or qualitative and quantitative data) is required to further investigate relevant questions (for example those around the relationships between social networking tools and project collaboration). These methods may involve gathering opinions from all stakeholders, especially from the teachers who use eTwinning for their professional learning.[vi]

In the context of professional learning, when investigating the value of LA for learners and their employers, we cannot make simple delimitations that equate types of systems with users and stakeholders and a given power relation. For example, an LA system that displays aggregates over learner behaviour in a Web analytics / business analytics way may primarily address site-operator users. However, these may act fully on behalf of the learners, the analytics system may have been co-designed with the help of learners, and the learners may

also profit as users when they see how the learning environment to which they contribute by their activities evolves over time. Conversely, an LA system that displays a range of personalized metacognition-supporting analytics may primarily target management users. The end users (learners) may not have had a say in the choice of the system, and it may prove to be most used by and useful for learners' managers, for evaluating employee performance. Therefore, ascertaining the potential of a system in supporting learning is not straightforward.

The case study illustrates that a great deal of professional learning occurs outside formal curricula, which concurs with many of the other studies of professional work and learning in this book. Professional learning can be self-organised and self-regulated, is not standardised and differs between learners and learner groups. The requirements for LA in these sorts of settings are highly context-dependent. The specific context (for example in professional learning) determines the target variables and success measures for goal attainment, which can vary from a group to one learner to another within the same cohort of learners. Given the multitude of users, stakeholders, use cases, and topics, there are important questions around who decides on the design, choice, and use of learning analytics for technology-enhanced professional learning.

One such question is the extent to which the use of LA is optional or mandatory. While LA may today seem like an additional, optional and 'fun' tool, this optionality is likely to change with the increasing maturation and professionalisation of the LA field, with LA becoming an integral component of "learning environments". Just as Web sites use instruments for measuring 'click-throughs' (a standardised measure) to earn money from the number of people who have viewed the site, LAs will likely be standardised and made non-optional in future learning environments. Selection of measures to use then may no longer be the choice

of the site operator. This development will exert pressure on these operators and/or teachers who work with the site to "teach to the test" – to design materials that will lead learners to exhibit the "right" behaviour.[vii] Likewise, learners may be asked to submit their "LA portfolios" in addition to or instead of other measures of learning outcomes, and therefore begin to "learn to the test (that is, the LA)". This vision appears not too far-fetched in view of the number of participants registering for Massive Online Open Courses (MOOCs) (Siemens, 2012), for which assessment of learning outcomes remains a major factor limiting growth. For example, Coursera[viii], a company partnering with universities in offering MOOCs, is considering selling learner/learning data to potential future employers (Young, 2012). This vision of LA does not align well with the highly personalised and fluid nature of professional work and learning laid out in the chapters in sections 1 and 2 of this book.

Lastly, many LAs have built-in feedback and awareness systems that could support learners' self-regulation activities. Designers of LA systems and interfaces should build on what is known from the literature about how feedback can be used to enhance professional learning (Hattie & Timperley, 2007; Boshuzien & van der Weil, this volume; Siadaty et al., this volume). In user evaluations, system designers should monitor whether learners know when and how to use feedback to support their learning tasks.

Outlook

In the future, it is critical to involve learners in decisions around LA: what goals to pursue when supporting professional learning; whether to use LA and which ones; what data to record and analyse; which interaction choices to use; which measures to compute and how to evaluate; and how to ensure that LA do not contribute to information overload. The

development of LAs that support learning requires these specific questions to be raised and discussed, in a participatory requirements analysis and system development process. In other words, ideally all stakeholders of LA should be actively involved in the design process in order to help ensure the LA system designed meets their needs and is usable.[ix] We believe that such participation is the key to opening up the potential of LA for professional learning – and other forms of learning. First, an understanding of the interests and concerns of different stakeholders may be used to improve the design of the platform by website and platform operators (or, in general, the providers and managers of the learning / LA environment). Second, it would also be interesting to allow users of the platform to reflect and comment on the results of design and use, stimulating further-reaching improvements.

One example of user feedback being utilised to improve systems design can be seen in relation to privacy, as a constraint on – or future feature of – LAs. LAs can be viewed as a form of 'surveillance' technology. Questions that relate to surveillance technologies should be discussed with all stakeholders affected by the technology. Critical questions include: do the benefits of the technology systems (for example, enhanced learning) offset the disadvantages (for example, choices and behaviours being tracked)? Does the surveillance have effects in and of itself (for example, are there inhibiting factors, such as the knowledge of being observed continuously leading to restrained dialogue on the part of the user)? What are the effects of a learner's /teacher's/ manager's actions on others while these actions are under surveillance? How are the interests of various people weighted and reconciled? (Gürses, 2010). The LA community is aware of these privacy issues (see, for example, Campbell, DeBlois, & Oblinger, 2007; Greller & Drachsler, 2012). Yet the community operates on the basis of an oversimplified assumption that privacy can be safeguarded by anonymising or access-controlling specific types of personal data (see Berendt, 2012, for an

extended discussion). This view extends widely beyond the LA domain. In work contexts, two additional factors are important: first, employers must respect legal restrictions on employee surveillance, and second, for both employers and employees, the personal and financial consequences of a breach of trust in employment relations are, in most cases, likely to be more significant than those in learning relationships of instructor-learner, teacher-pupil, or company-customers.[x]

In general, LA's greatest potential in supporting informal, social professional learning lies in tools that learners themselves interact with. Social Semantic Web (SSW) tools, such as those described by Siadaty et al. (this volume), are particularly interesting. These tools take into account the diversity of real-world online tools that people use, particularly in informal learning settings. LA could be an interesting "piggyback" addition alongside these sorts of tools, and their addition appears feasible given that the heterogeneous data are already recorded and semantically analysed and transformed by the Semantic Technologies used for SSW tools. Siadaty et al. (this volume) offer a glimpse of the possibilities afforded by LA when used in tandem with SSW tools. A glimpse of the possibilities of LA additions to SSW tools is offered by the divergence between professional learners' self-reported attitudes and behaviour reported by Siadaty et al. (this volume): Learners, when describing themselves, usually stated that the organizational context influenced their setting of their learning goals. However, the data gathered during their learning activities showed that they relied just as much on their social context for setting their learning goals. An analysis of the self-report data and the behavioural data can surface these sorts of mis-alignments. Reflecting these sorts of discrepancies to learners could help them reflect on their self-regulated professional learning, leading them to new, productive insights.

Another direction for future research that brings together opportunities and challenges described in this paper is the use of LA in blended learning. The case study illustrates that learning rarely takes place within a single environment. Combinations of different learning environments – both digital and physical – are likely to increasingly become complex in continuing professional learning. A question arises as to what extent LA could – and should – span more (or even all) these environments, whether 'online' or in 'off-line', physical settings. Technically, research in this domain could draw on methods from Web analytics and Web mining to collect and analyse data from different communication and distribution 'channels' between businesses and customers (Teltzrow & Berendt, 2003). While this could yield interesting insights into the use of online versus face-to-face activities (Brian McNely, Gestwicki, Holden Hill, Parli-Horne, & Johnson, 2012), the extended data collection may require too much surveillance and could cause privacy problems. Therefore, new methods for empowering users to take part in or opt out of analytics, and to make informed choices around analytics, will become critical. The field of LA has a unique opportunity to mature by embracing these novel and difficult challenges through participatory design that truly reflects the concerns of all the different stakeholders affected.

References

Berendt, B. (2012). More than modelling and hiding: towards a comprehensive view of Web mining and privacy. *Data Mining and Knowledge Discovery*, *24*(3), 697-737.

Berendt, B., Clarke, D., De Wolf, R., Gao, B., Peetz, T., Pierson, J., Preibusch, S., & Sayaf, R. (2012). *SPION Deliverable 5.1. Report on Research Activities (Feedback and Awareness*

Tools). COSIC Technical Report, K.U. Leuven, Belgium. Retrieved from http://www.cosic.esat.kuleuven.be/publications/article-2302.pdf

Berlanga, A. J., & Vuorikari, R. (Eds.). (2012). Symposium – Learning Networks for Professional Development: Current Research Approaches and Future Trends. In *Proceedings of the 8th International Conference on Networked Learning 2012.* Maastricht, The Netherlands: Springer.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief.* Brief submitted to the US Department of Education. Retrieved from http://www.ed.gov/edblogs/technology/files/2012/03/edm-la-brief.pdf

Cambridge, D., & Perez-Lopez, K. (2012). First steps towards a social learning analytics for online communities of practice for educators. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)* (pp. 69-72). New York: ACM.

Campbell, J.P., DeBlois, P.D., & Oblinger, D.G. (2007). Academic Analytics: A new tool for a new era. *EDUCAUSE Review*, *42*(2), 41-57.

Cao, Y., Klamma, R., Pham, M., & Vuorikari, R. (2012). Social Network Analysis Methods for Teacher Networks. In R. Vuorikari et al. (Eds.) (2012). *Teacher Networks - Today's and tomorrow's challenges and opportunities for the teaching profession. European Schoolnet* (pp. 37-51). Retrieved from http://service.eun.org/teachers-newsletter/TellNet_Teacher_Networks_web.pdf

Ferguson, R. (2012). *The State of Learning Analytics in 2012: A Review and Future Challenges.* Knowledge Media Institute: Technical Report KMI-12-01. Retrieved from http://kmi.open.ac.uk/publications/pdf/kmi-12-01.pdf

Ferguson, R. & Buckingham Shum, S. (2012). Social learning analytics: Five approaches. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)* (pp. 23-33). New York: ACM.

Greller, W. & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Educational Technology & Society*, *15*(3), 42-57.

Gürses, S. (2010). *Multilateral Privacy Requirements Analysis in Online Social Network Services.* PhD thesis, K.U. Leuven, Dept. of Computer Science. Retrieved from http://www.cosic.esat.kuleuven.be/publications/thesis-177.pdf

Hattie, J. & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*, 81-112.

Haythornthwaite, C. (2001). The strength and the impact of new media. In *Proceedings of the 34th Annual Hawaii International Conference On System Sciences.* Retrieved from http://computer.org/proceedings/hicss/0981/Volume%5C%201/09811019abs.htm

Littlejohn, A., Milligan, C., & Margaryan, A. (2012). Charting collective knowledge: supporting self-regulated learning in the workplace. *Journal of Workplace Learning*, *24*(3), 226-238.

McAuley, J., O'Connor, A., & Lewis, D. (2012). Exploring reflection in online communities. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)* (pp. 102-110). New York: ACM.

McNely, B.J., Gestwicki, P., Holden Hill, J., Parli-Horne, P., & Johnson, E. (2012). Learning analytics for collaborative writing: a prototype and case study. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)* (pp. 222-225). New York: ACM.

OECD (2009). *Creating Effective Teaching and Learning Environments: First results from TALIS.* Paris. Retrieved from www.oecd.org/edu/school/43023606.pdf

Pham, M. C., Cao, Y., Petrushyna, Z., & Klamma, R. (2012). Learning Analytics in a Teachers' Social Network. In *Proceedings of the Eighth International Conference on Networked Learning (NLC 2012), Maastricht, the Netherlands, April 2-4, 2012* (pp. 414 - 421).

Santos, J.L., Govaerts, S., Verbert, K., & Duval, E. (2012). Goal-oriented visualizations of activity tracking: a case study with engineering students. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)* (pp. 143-152). New York: ACM.

Schlager, M. S., Farooq, U., Fusco, J., Schank, P., & Dwyer, N. (2009). Analyzing Online Teacher Networks: Cyber Networks Require Cyber Research Tools. *Journal of Teacher Education*, *60*(1), 86-100.

Siadaty, M., Gašević, D., Jovanović, J., Milikić, N., Jeremić, Z., Ali, L., Giljanović, A., & Hatala, M. (2012). Learn-B: a social analytics-enabled tool for self-regulated workplace learning. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)* (pp. 115-119). New York: ACM.

Siemens, G. (2012). *MOOCs are really a platform.* Elearnspace. Retrieved from http://www.elearnspace.org/blog/2012/07/25/moocs-are-really-a-platform/

Siemens, G. & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12)* (pp. 252-254). New York: ACM.

Siemens, G., & Gasevic, D. (2012). Learning and Knowledge Analytics. *Journal of Educational Technology & Society*, *15*(3), 1–2.

Sloep, P., & Berlanga, A. (2011). Learning Networks, Networked Learning. *Comunicar*, *19*(37), 55–64.

Song, E., Petrushyna, Z., Cao, Y., & Klamma, R. (2011). Learning Analytics at Large: The Lifelong Learning Network of 160,000 European Teachers. In C. D. Kloos, D. Gillet, R. M. Crespo García, F. Wild, & M. Wolpers (Eds.), *Towards Ubiquitous Learning* (pp. 398–411). Berlin, Heidelberg: Springer Berlin Heidelberg. LNCS 6964.

Suthers, D. D., Dwyer, N., Medina, R., & Vatrapu, R. (2010). A framework for conceptualizing, representing, and analyzing distributed interaction. *International Journal of Computer-Supported Collaborative Learning*, *5*(1), 5-42.

Teltzrow, M., & Berendt, B. (2003). Web-Usage-Based Success Metrics for Multi-Channel Businesses. In *Proceedings of the WebKDD 2003 Workshop - Webmining as a Premise to Effective and Intelligent Web Applications. August 27th, 2003, Washington DC, USA. Held in conjunction with SIGKDD 2003* (pp. 17-27). Retrieved from http://people.cs.kuleuven.be/~bettina.berendt/Papers/teltzrow_berendt_webkdd03.pdf

van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). Analytics in higher education: Establishing a common language. *Educause Learning Initiative*, *1*, 1-11.

Vuorikari, R., Garoia, V., Punie, Y., Cachia, R., Redecker, C., Cao, Y., Klamma, R., Pham, M.C., Rajagopal, K., Fetter, S., & Sloep, P. (Eds.) (2012). *Teacher Networks - Today's and tomorrow's challenges and opportunities for the teaching profession. European Schoolnet.* Retrieved from http://service.eun.org/teachers-newsletter/TellNet_Teacher_Networks_web.pdf

Vuorikari, R., & Scimeca, S. (2013). Social Learning Analytics to study Teachers' Large-scale Professional Networks. *In Open and Social Technologies for Networked Learning* (pp. 25-34). Berlin, Heidelberg: Springer, IFIP Advances in Information and Communication Technology, vol. 395.

Wastiau, P., Blamire, R., Kearney, C., Quittre, V., Van de Gaer, E., & Monseur, C. (2013). The Use of ICT in Education: a survey of schools in Europe. *European Journal of Education*, *48*(1), 11–27.

Young, J. R. (2012, July 19). Inside the Coursera Contract: how an upstart company might profit from free courses. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/How-an-Upstart-Company-Might/133065/

Notes

[i] Records of requested materials (such as clicked-on Web pages) and user input (such as queries or other typed-in data)

[ii] Classifiers are models in machine learning. They include rules for classifying or predicting whether an individual belongs to a certain class. For example, a classifier might predict that someone with certain traits and behaviour is likely to obtain high scores in a test, or is likely to not finish a certain activity.

[iii] http://www.etwinning.net

[iv] http://tappedin.org

[v] webarchive.nationalarchives.gov.uk/*/http:/www.teachernet.gov.uk/

[vi] Other stakeholders include the site providers, school managements, school authorities, pupils, parents, ... Depending on the specific questions and value decisions, the voices of stakeholder groups will be heard and/or will influence interprations and design decisions. We cannot cover all these decisions in detail here, but want to focus on the teachers as stakeholders.

[vii] Greller and Drachsler (2012) go beyond this by claiming that „the real dangers [are] that the extended and organized collection of learner data may not so much bring added benefits to the individual, but instead [provide] a tool for HEIs, companies, or governments to increase manipulative control over students, employees, and citizens, thereby abusing LA as a means to reinforce segregation, peer pressure and conformism rather than to help construct a needs-driven learning society." (p. 54).

[viii] htttp://www.coursera.org

[ix] Agile development methodologies such as those proposed by Clow (this volume) may be a solution..

[x] See http://epic.org/privacy/workplace/ for an extensive resource collection.

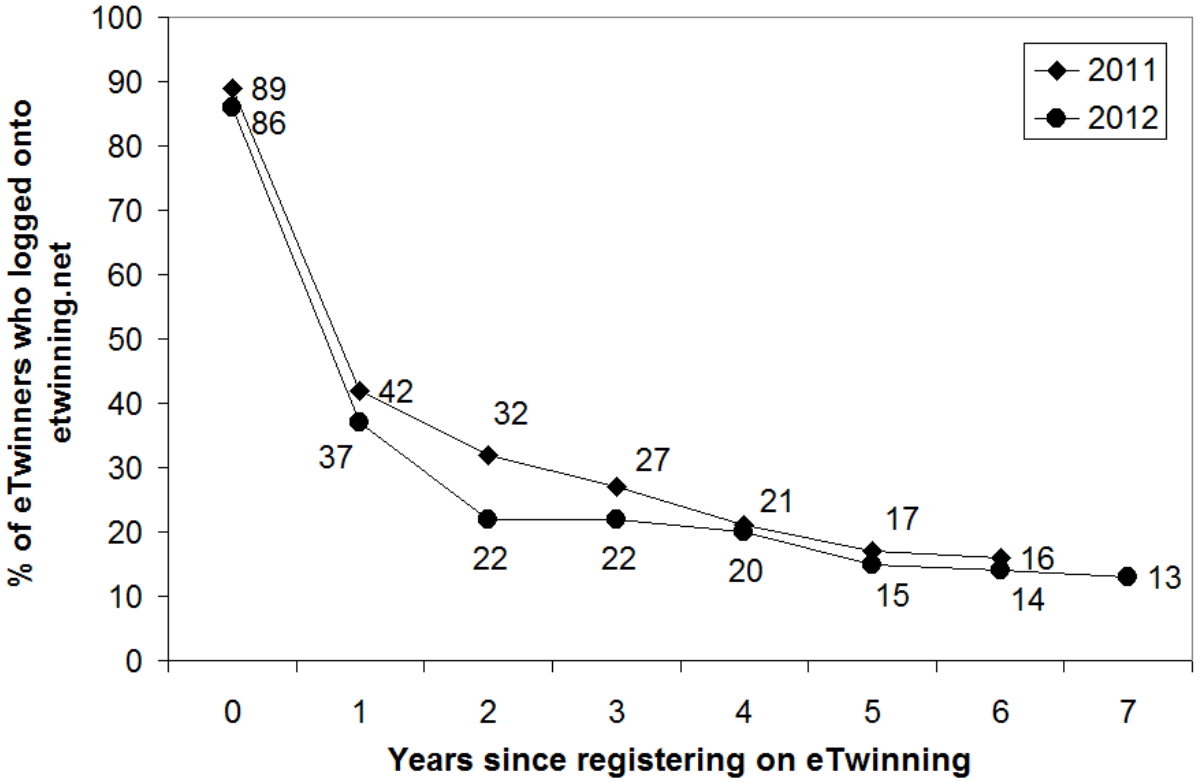


Fig. 13.1. eTwinning retention rate in 2011. Data on returning eTwinners by the year of their registration: "0 year" refers to eTwinners who registered in 2011; "1 year" = in 2010, etc.

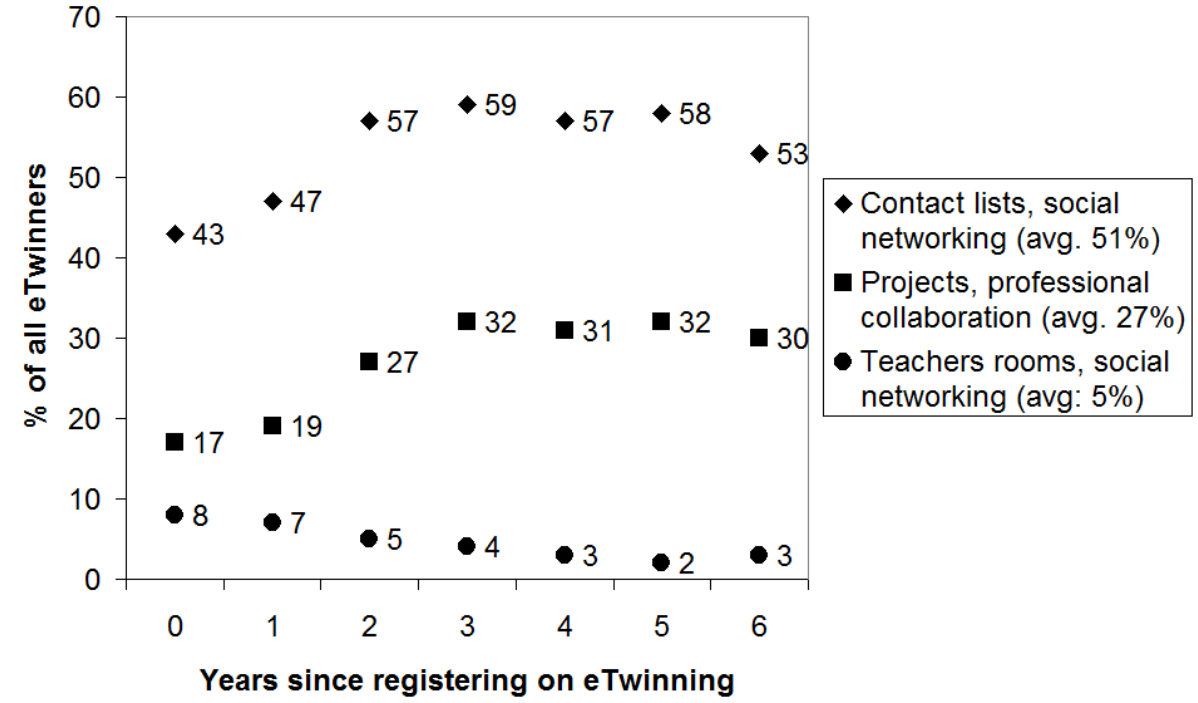


Fig. 13.2. eTwinners engagement on the portal disaggregated by the year of registration. "0 year" refers to eTwinners who registered in 2011; "1 year" to those registered in 2010, etc.

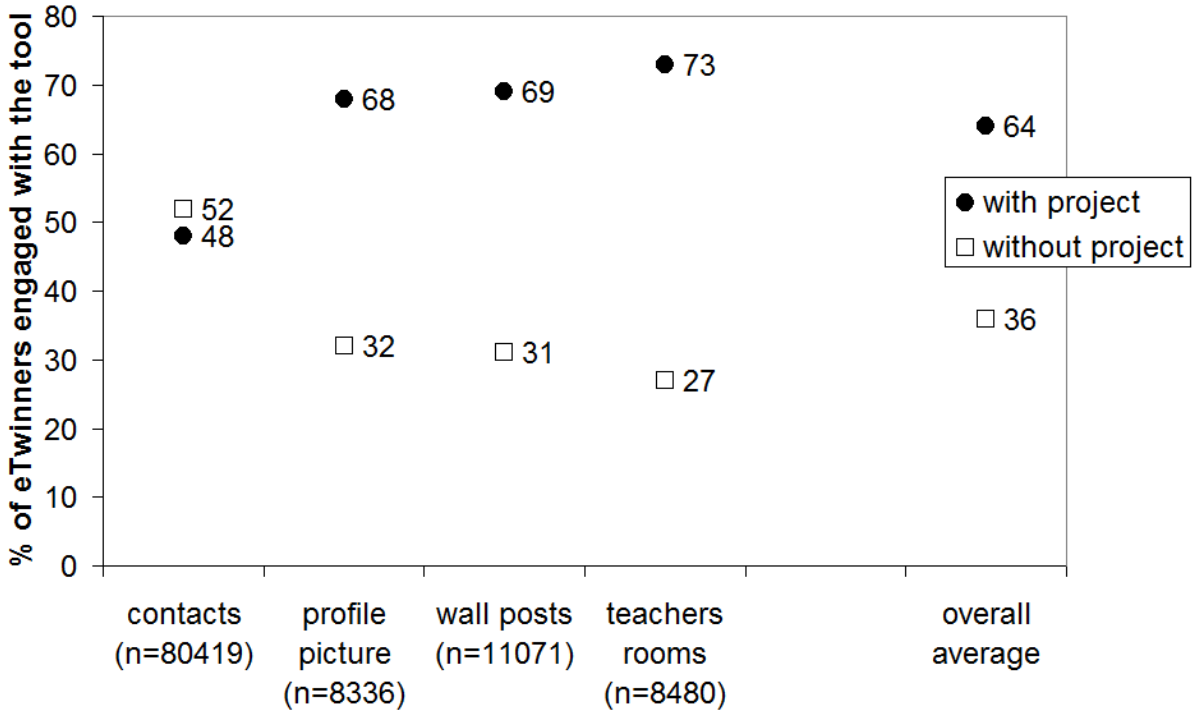


Fig. 13.3. eTwinners' use of social networking tools, divided by eTwinners with projects and without. Snapshot of data extracted from SteerCom-Desktop tool (Feb 2012).

8

PAPER 6:

“Tool Clinics" - Embracing multiple perspectives in privacy research and privacy-sensitive design

- What are the forums for encouraging collective action in participatory sensing? Can we encourage system designers to consider social justice during design by framing design as a collective action problem? Can participatory sensing open new avenues for consumers and citizens to organize collective action?
- Could sensing data help us "diagnose" people's moral predispositions? (And therefore political behavior?)
- What factors in sorting and categorization processes make people feel that resulting algorithmic treatment is fair or unfair?

References

1 Laura Hueya, Kevin Walby, and Aaron Doyle. *Surveillance and security: technological politics and power in everyday life*, chapter Cop watching in the downtown eastside: exploring the use of (counter) surveillance as a tool of resistance, pages 149–165. Routledge, 2006.
2 Marie C. Oetzel and Sarah Spiekermann. Systematic methodology for privacy impact assessments. *European Journal of Information Systems*, July 2013.
3 John Rawls. *A theory of justice*. Cambridge, MA: Harvard University Press, 1999.
4 Amartya Sen. *The idea of justice*. Cambridge, MA: Harvard University Press, 2009.

4.3 "Tool Clinics" – Embracing Multiple Perspectives in Privacy Research and Privacy-Sensitive Design

Anthony Morton (University College London, GB)
Bettina Berendt (KU Leuven, BE)
Seda Gürses (KU Leuven, BE)
Jo Pierson (Free University of Brussels, BE)

4.3.1 Focalism – The Challenge

Computer scientists or engineers are continually asked to "solve problems" or "improve" existing situations, by selecting from available design features to produce the "best" technical solution. For example, a software developer faced with the problem of securing data must choose between different encryption algorithms – each with different characteristics. Factors such as strength of encryption, speed of encryption, usability, key management and hardware requirements must all be considered. Other requirements such as the sensitivity and amount of data to be protected, the estimated resources of potential attackers, the operational context of the required solution, etc. must also be taken into account. It is impossible for any solution to be 100% perfect, e.g. encrypting data with no detectable delay using an algorithm which cannot be broken. Trade-offs during the design and development process are therefore inevitable as requirements are balanced, e.g. speed vs. strength of encryption. These trade-offs are dilemmas faced by the specialist in arriving at the final design. However, what is the "best solution", and who decides what "best" means, requires more involved discussion and reflection. The engineer, with their narrow focus on solving the technical problem, might not be best equipped to solely decide what the optimum solution is, particularly if there are likely to be unintended consequences when the solution is deployed, or the proposed technology is decoded differently by users, those directly or indirectly affected, and other stakeholders.

The desire of specialists – particularly those in the fields of science or technology – to frame complex and messy situations as a single problem to be solved by technology – for which only they have the answer – often leads to overconfidence in the envisaged solution, an overemphasis on intended consequences, and a tendency to focus narrowly on one or a few aspects of the problem. This is typically identified as a form of "technological determinism", a perspective which consists of two parts: (1) the belief that technological developments take place outside society, independently of social, cultural, economic and political forces; and (2) the assumption that technological change causes or determines social change [31]. This kind of technologically deterministic approach can result in bigger problems than the one originally being solved because the understanding of the original problem situation was incomplete or wrong; Tenner [26] calls these the "unintended consequences" of technological innovation, e.g. the increasing resistance of certain strains of bacteria to antibiotics. However, unintended consequences are not restricted to technological innovation, but occur in political science, organisations, medicine and public health, ecology and social systems [11, 26]. Ehrlinger and Eibach [11] observe:

> "*[F]ocalism, or a tendency to focus narrowly on one or a few variables, [. . .] with respect to the intended consequence can result in a neglect of important information regarding alternative, unintended consequences – including information that is knowable and plainly relevant to predictions*" (p. 60)

Using a computer simulation, Ehrlinger and Eibach [11] showed that participants who were "defocused" by being encouraged to consider a wider system of variables, tended to make more accurate predictions and were less optimistic about the proposed solution. This suggests that viewing problems more holistically – particularly from multiple perspectives – can improve decision-making and increase the chances of successful technology development. Focalism – probably first suggested by Wilson et al. [30] – is essentially the same as "focusing illusion" proposed by Schkade and Kahneman [21] and Loewenstein and Schkade [14]. They found that when people are asked to predict their emotive reaction to a major event (e.g. the loss of employment), they typically concentrate on their likely responses to the focal event, to the exclusion of possible effects of other non-focal events (e.g. new opportunities to start a business or retrain). A practical example of people's tendency to ignore other events when their attention is focused elsewhere – inattentional blindness – is described the study by Simons and Chabris [23] in which most people missed a gorilla appearing during a video, when asked to concentrate on the number of times the ball was passed between particular basketball players.

We propose that the notion of focalism is equally applicable to scientists and technologists, who are often reluctant to challenge assumptions surrounding a problem, and principally concentrate on finding a solution to the problem as they perceive it, without adequate consideration of: (1) what it is that actually needs to be achieved – not from only one viewpoint; (2) any foreseeable consequences of the proposed solution; (3) and the viewpoints of other affected and/or interested actors who may have different priorities. We suggest this can be viewed as "solution focalism", and we propose that de-focusing may best be achieved by making other viewpoints salient As Genus observes, "*the employment of participatory approaches has been proposed to accommodate the interests of a wide range of actors holding different value positions, while minimising the potential risks associated with technology development.*" [12]

The problems of focalism are not restricted to technology development. It also reduces the efficacy of privacy research and privacy-sensitive design. For example, Privacy Enhancing

Technologies (PETs), such as Privacy Bird and Privacy Finder[2], appeared *prima facie* at the time to offer useful technical solutions to the problem of managing people's privacy. Both PETS use a protocol published in 2002 by the Platform for Privacy Preferences Project (P3P) [7] that enables web sites and applications to describe their privacy policy in XML. However, they have failed to become widely accepted and deployed. In 2003, the adoption rate of P3P was broadly flat at around 10% [10], partially due to the limited functionality of the first P3P user agents, and user interface problems [8]. Reay et al [18] observed that "*P3P adoption has stagnated in a niche position; it appears that browser implementers simply do not have enough market incentive to expend the resources needed to develop and integrate P3P 1.1 user agents*" (p. 162). Those browser implementers that did implement P3P made such fundamental technical mistakes that P3P was easily circumvented by publishing invalid policies [9]. Companies who chose not to use P3P suffered no consequences, which underlined the fact that P3P – albeit an elegant technical design – also required, as a minimum, enforcement external to itself, either through government regulation or industry self-regulation, both of which never materialised. The development of P3P may have benefited from collaborative design and development informed by a critical assessment of the perspectives of browser developers, the interests and technical capabilities of those who host and manage web sites, and the role of regulators. Certainly, there is much to be learned from the P3P experience that can be used to look at contemporary proposals for privacy-sensitive design. Focalism has also influenced the empirical aspects of privacy research. Many privacy studies have focused on the user experience with different interfaces and privacy controls, without thinking more holistically and considering the context in which the tool is used, the primary goals the user is trying to achieve, or the interaction of these goals with the interests of other affected stakeholders.

We propose a "tool clinic" to encourage a collaborative (re)consideration of a technological solution, research technique or other artefact, in order to critically assess its design, development and deployment from multiple perspectives. Another objective is to turn such solutions or artefacts into a tool for exploring the problem space. For example, what is the privacy problem when we look at it through a solution such as P3P? Finally, a tool clinic can be used to provide those who are developing the solutions with a setting to rethink the framing and presentation of their solutions. The term "tool clinic" emphasizes the motivation for embarking on this exercise. Athletes dedicated to improving some specific skill routinely go to a "rebound clinic" (in basketball) or a "dribbling clinic" (in football). The use of the word "clinic" does not indicate that a tool clinic provides a specific fix for problems, best practice guidelines, or solution templates – a typical panacea sought by those in the field of engineering. Rather, a tool clinic provides a framework and approach for multiple-perspective formative exploration and review of a technological solution, research technique or other artefact under development. The objective is to reflect from different perspectives on practices around the development, encoding, use, domestication, decoding and sustainability of a tool to gain quasi-ecological validation. In this sense, a tool clinic is more like a "law clinic", where law students study law and practice the adversarial legal process in context, or "design crits", during which designers learn to critique and receive critique of their work from others in the arts, academia or design practice.

[2] Privacy Bird was initially developed by AT&T. Privacy Bird and Privacy Finder are managed by Carnegie Mellon University's Usable Privacy and Security Laboratory.

4.3.2 Existing Uses of Multi-perspective Formative Exploration and Review

It is important to demonstrate that similar approaches to the suggested "tool clinic" are already used successfully in areas of industry and academia. This section describes some existing techniques that use a multi-perspective and collaborative approach.

In industry, disaster recovery practitioners often use corporate "war games" – a term originating from the military – to simulate a potential disaster situation (e.g. the loss of a data centre), and step through its disaster recovery plans to ensure they operate correctly. This avoids situations such as employees not being able to relocate to a cold-standby office building due to keys or swipe-cards not being readily available because the security department was excluded from disaster recovery planning. The use of disaster recovery simulations involving all affected areas of the business ensures disaster recovery plans are considered from multiple perspectives. A related technique to war games, the "Red Team"[3] review, also originated in the military as a means of assessing plans in an operational context from the perspectives of adversaries, affected areas of the military and their partners. Like war games, a Red Team review subjects a problem, plan, process, technique or artefact (e.g. tool, document, service, software product, etc.) to rigorous scrutiny by trained team members and experts. One of the authors of this report has been involved in Red Team reviews of complex commercial bid documents by the technical design and implementation, financial, service management and legal areas of a business organisation.

Gaining multiple perspectives is a technique also used by Soft Systems Methodology (SSM), which emerged in the 1980s from Checkland's work [5, 6]. SSM is a framework for organising the exploration of messy, complex problems as a learning *system*, and therefore failures in projects, processes etc. are viewed as a *systems failure*. Checkland [5] suggests that to fully understand a system it is necessary to consider its purpose from different viewpoints. This systemic pluralism represents one aspect of the "soft" systems approach, which aims to construct a rich picture of a problem, encompassing different viewpoints, rather than the reductionist focus of systems engineering. These different viewpoints, or *Weltanschauungen*, represent unquestioned models of the world that makes the system meaningful for study [5, 6]. It is important to stress that although SSM views problems as a *system*, it is not a representational model of reality; it is epistemological, not ontological; just because SSM views a situation *as if it were* a system, does not mean *it is* a system [6], e.g. a computer system.

To facilitate understanding of the reasons for failures, Checkland created the idea of a *formal system model* (FSM), which is a "*general model of any human activity system*" [5]. Comparison between the formal system model and the conceptual model of the problem situation under investigation is an intrinsic part of the SSM process, as it identifies flaws, weaknesses and omissions in the conceptual model, facilitating its improvement. The improved conceptual model can be compared with the real-world situation to determine which desirable or feasible changes are required [5, 6]. A project specific form of the FSM has been developed by Fortune et al [28] for use in analysing project failures, such as large-scale building projects [29].

The existing multi-perspective techniques described thus far, not only subject items to rigorous review, but encourage collaborative improvement and design. Soliciting the

[3] A "Red Team" is defined as "a team that is formed with the objective of subjecting an organisation's plans, programmes, ideas and assumptions to rigorous analysis and challenge. Red teaming is the work performed by the red team in identifying and assessing, inter alia, assumptions, alternative options, vulnerabilities, limitations and risks for that organisation." [1].

viewpoints of stakeholders, potential users of a technology or service, and those affected by it, can dramatically improve its quality. The notion of collaborative development and improvement to ensure effort is not expended on features or services that customers do not require, is key to the notion of "*the lean startup*" [19] used by many Internet companies. The lean startup philosophy suggests that companies release a "minimum viable product" – a "*version of a new product which allows a team to collect the maximum amount of validated learning about customers with the least effort*" [19] – to a subset of sympathetic customers, such as early adopters. The release of a minimum viable product is part of an iterative prototyping process, collecting suggestions for improvement, learning how customers use the product and what they want from it. The use of minimum viable products allows business to understand how customers actually decode the technology or service being provided; the product must be viable in that the customer must value what it provides. Use of minimum viable products should be an iterative learning process, generating ideas and collecting data about product use.

One existing approach to answer the question posited earlier, "*Who decides what 'best/better' really means?*" is constructive technology assessment (CTA). The latter fits within the long-standing tradition of Science and Technology Studies (STS), which investigates how the things that it studies are being constructed. The STS domain has increased its scope over the years, starting with scientific knowledge and expanding to artefacts, methods, materials, observations, phenomena, classifications, institutions, interests, histories, and cultures [24]. One of the most prominent ways to apply the thinking in STS in the real world has been the CTA approach. The objective of CTA is to "*produce better technology in a better society*" [12] by taking a more social constructionist position, and moving "*beyond technological determinism towards an evolutionary view of technology development*" [12]. This is done by advising on interventions in early stages of technology development based on the assessment of possible problems and risks that these technologies could pose for society [25]. CTA emphasises the importance of including a wide range of actors to anticipate the potential impact of a technological development ("vermaatschappelijking" of technology [27]) and decide on improvements to it, thus facilitating social learning. It should be stressed that CTA is not a research method, but an overall approach into which participatory techniques may be placed. Genus [12] suggests moving away from the interventionist and prescriptive stance of existing CTA approaches towards a more discursive, democratic and reflective process because "*contention and openness to criticism are prerequisites for producing reflective socio-technical expertise*" [12]. This is also known as "participatory technology assessment" [13]. The use of a modified form of CTA to address the ethical problems caused by technology is proposed by Palm and Hansson [16] as part of a continuous dialogue between developers and affected actors. For emerging technologies, Merkerk and Smith [27] propose a three-step CTA approach, using permutated dialogue workshops attended by insiders and outsiders to the item under review to consider selected issues about the proposed technology and reflect on different technology scenarios.

In order to apply multi-perspective formative exploration and review of technological solutions or tools in early stages of development, different types of multi-method approaches have been developed. One of the most elaborate ones is the living laboratory approach. The 'living lab' is a specific type of test and experimentation platform (TEP), which refers to facilities and environments for (joint) innovation including testing, prototyping and confronting technology with usage situations [3]. Living labs are facilities for designing, developing, testing and evaluating communication technologies and services in early stages of the innovation process. They do so by involving (early) users, in line with the CTA

perspective. However they can also be configured as open and innovation-oriented platforms that involve various technology experts, disciplines and/or stakeholders in different stages of technology design, development and testing [17]. Thus, we discern three main ways to put living labs[4] into action as: (1) a platform for open innovation; (2) a user-driven research methodology; and (3) an experimental setting [20].

4.3.3 Perceived Research Gap in Privacy

Most privacy researchers agree that privacy is contextual and dependent upon information use, information sensitivity and the trust in the entity collecting, storing, processing and disseminating the information entrusted to it [2]. Furthermore, users engaged in technology mediated interactions with other parties will have expectations and assumptions about the technology, the providing organisation and other partners in communication [2]. If these assumptions and expectations are violated, the user is likely to have an emotional reaction and reject the technology and/or providing organisation [2]. A practical example of this was the launch of Google Buzz. Gmail users believed they were only signing onto Gmail as usual, when they were actually being enrolled in Google Buzz [22]. It would appear the developers of Buzz did not take into account: (1) that people's primary task was to access their e-mail and hence they would likely "swat away" any dialogue boxes without properly reading them; and (2) that people's mental model is that Gmail is a tool to access their e-mail and not a social networking service.

User studies may aid developers and designers in foreseeing likely troubles that users may have with a given design. However, the task of achieving an understanding of the complexity of the privacy problem, and translations of this problem into the technical solution space may benefit greatly from a multi-perspective approach. This is line with the notion of *contextual integrity* (CI) by Nissenbaum [15], which is used to answer whether a situation contained a privacy breach or not. CI is guided by norms of appropriateness (i.e. norms that govern what can be disclosed in a certain context or situation) and norms of distribution (i.e. norms which assess the transfer of personal information from one party or context to another context). This demonstrates how not all publicly revealed information or information collected in the public space, is meant for every form of public use. "*Just because something is publicly accessible does not mean that people want it to be publicized. Making something that is public more public is a violation of privacy.*" [4]

Addressing the privacy implications of increasingly complex, powerful and ubiquitous computing will be even more of a challenge than Buzz, as the potential for unintended consequences is even greater than before. However, privacy researchers and practitioners continue to work largely in isolation, concentrating on people's use of different user interfaces for privacy control, and have largely ignored existing cross-disciplinary collaboration techniques such as those described above.

4.3.4 Future Directions for Researchers and Practitioners

Tool clinics are essentially practices, and they need to be living practices – thus future directions are not only researching, but also must be *doing* tool clinics. We have performed a first *ad hoc* requirements analysis for tool clinics at the Dagstuhl Seminar itself (i.e. we "clinicked"

[4] In Europe living labs are associated in the European Network of Living Labs (ENoLL) which was set up under the auspices of the Finnish EU presidency in 2006 and since the 6th wave of call for new members in March 2012 consists of over 300 accepted members.

the tool clinic idea) and have seen the challenges the concept poses. Most importantly, our clinic participants expressed concerns about exposing their methods, approaches and original ideas to a critical audience. Further issues were raised with respect to matters of intellectual property. Some of these problems are likely to stem from the employment requirements and the working conditions of senior and junior researchers. They also often associated the word "clinic" with doctoring their (software) artefacts with others, a goal that we only partially share.

Based on this experience, our next step will be to develop a tool clinic as a new event format for a scientific conference, ideally at a renowned computer-science conference. This will combine the tool-centric nature of a demo session, the protected space of work-in-progress afforded by a workshop, and the mentoring spirit of a doctoral workshop[5].

The format of a tool clinic session could typically consist of three steps (inspired by the CTA and Privacy by Design approach):

1. Identifying particular affordances of the technological solution, research technique or other artefact and possible (unintended) consequences for people and society;
2. Gathering perspectives and practices of different experts, disciplines and/or stakeholders (e.g. users, policy makers, industry, etc.) linked with the development, deployment and sustainable evolution of a particular tool, solution, technique or artefact;
3. Informing and advising on technological design of the tool or solution, in order to avoid negative consequences and to further positive outcomes.

We foresee three essentially needed incentives for participation: (1) enlisting big names in the field who can signal through their own example that "grown-ups too can learn"; (2) a broad-enough team of participants to represent a wide range of perspectives; and (3) a follow-up that makes it worthwhile to put oneself into the ring. For the first two, we can draw on our respective scientific networks. A special issue in a good journal is one option for creating the third incentive, and further developments of the tool clinic method described in the introductory article of this special issue are among the next intended research activities.

4.3.5 Acknowledgement

We acknowledge support from the Strategic Basic Research (SBO) Programme of the Flemish Agency for Innovation through Science and Technology (IWT) in the context of the SPION project[6] under grant agreement number 100048.

References

1 Red Teaming Guide. Technical report, UK Ministry of Defence (2nd Edition), 2013.
2 Anne Adams and Martina Angela Sasse. Privacy in multimedia communications: Protecting users, not just data. In *People and Computers XV – Interaction without Frontiers*, pages 49–64, 2001.
3 Pieter Ballon, Jo Pierson, and Simon Delaere. *Designing for Networked Communications: Strategies and Development*, chapter Fostering Innovation in Networked Communications: Test and Experimentation, pages 137–166. IGI Global, 2007.
4 Danah Boyd. Making sense of privacy and publicity. Technical Report MSR-TR-2010-25, Microsoft Research, 2010.

[5] In this way the tool clinic approach has some resemblance with a 'crit' as done in art schools. This is a critique session, in which a student's artwork is formally presented to and evaluated by a group of faculty and peers, responding with feedback: comments, questions, advice, cheers, jeers, and tears.
[6] www.spion.me

5 Peter Checkland. *Systems thinking, systems practice*. John Wiley & Sons Ltd., 1981.
6 Peter B. Checkland and Jim Scholes. *Soft Systems Methodology in Action*. John Wiley & Sons Ltd., 1999.
7 Lorrie Faith Cranor. *Web privacy with P3P - the platform for privacy preferences*. O'Reilly, 2002.
8 Lorrie Faith Cranor. P3P: Making Privacy Policies More Useful. *IEEE Security and Privacy*, 1(6):50–55, November 2003.
9 Lorrie Faith Cranor. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *Journal of Telecommunications and High Technology Law*, 10(2), December 2012.
10 Lorrie Faith Cranor, Simon Byers, and David Kormann. An Analysis of P3P Deployment on Commercial, Government, and Children's Web Sites as of May 2003. Technical report, AT&T Labs-Research, 2003.
11 Joyce Ehrlingera and Richard P. Eibachb. Focalism and the failure to foresee unintended consequences. *Basic and Applied Social Psychology*, 33(1):59–68, 2011.
12 Audley Genus. Rethinking constructive technology assessment as democratic, reflective, discourse. *Technological Forecasting and Social Change*, 73(1):13–26, 2006.
13 Simon Joss and Sergio Bellucci, editors. *Participatory technology assessment: European perspectives*. Centre for the Study of Democracy, University of Westminster, 2002.
14 George Loewenstein and David Schkade. Wouldn't it be nice? predicting future feelings. *Well-being: The foundations of hedonic psychology*, pages 85–105, 1999.
15 H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.
16 Elin Palm and Sven Ove Hansson. The case for ethical technology assessment (eTA). *Technological Forecasting and Social Change*, 73(5):543–558, 2006.
17 Jo Pierson and Bram Lievens. Configuring living labs for a 'thick' understanding of innovation. *Ethnographic Praxis in Industry Conference Proceedings*, 2005(1):114–127, 2005.
18 Ian K. Reay, Patricia Beatty, Scott Dick, and James Miller. A Survey and Analysis of the P3P Protocol's Agents, Adoption, Maintenance, and Future. *IEEE Transactions on Dependable and Secure Computing*, 4(2):151–164, April 2007.
19 Eric Ries. *The Lean Startup: How Constant Innovation Creates Radically Successful Businesses*. Penguin Books Limited, 2011.
20 S.C. Sauer. *User innovativeness in living laboratories: everyday user improvisations with ICTs as a source of innovation*. PhD thesis, Universiteit Twente, Enschede, September 2013.
21 David Schkade and Daniel Kahneman. Does living in california make people happy? a focusing illusion in judgments of life satisfaction. *Psychological Science*, 9(5):340–346, September 1998.
22 Maggie Shiels. Google buzz 'breaks privacy laws' says watchdog. BBC News, February 2010.
23 Daniel J. Simons and Christopher F. Chabris. Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074, 1999.
24 Sergio Sismondo. *The handbook of science and technology studies*, chapter Science and Technology Studies and an Engaged Program. The MIT Press, 2008.
25 Wim A. Smit and Ellen C.J. van Oost. *De wederzijdse beïnvloeding van technologie en maatschappij: een Technology Assessment-benadering*. Bussum: Coutinho, 1999.
26 Edward Tenner. *Why things bite back: technology and the revenge of unintended consequences*. Vintage Books, 1997.
27 Rutger O. van Merkerk and Ruud E.H.M. Smits. Tailoring CTA for emerging technologies. *Technological Forecasting and Social Change*, 75(3):312–333, 2008.

28 Diana White and Joyce Fortune. The project-specific formal system model. *International Journal of Managing Projects in Business*, 2(1):36–52, 2009.

29 Diana White and Joyce Fortune. Using systems thinking to evaluate a major project: The case of the gateshead millennium bridge. *Engineering, Construction and Architectural Management*, 19(2):205–228, 2012.

30 Timothy D. Wilson, Thalia Wheatley, Jonathan M. Meyers, Daniel T. Gilbert, and Danny Axsom. Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 78(5):821–836, 2000.

31 Sally Wyatt. Technological determinism is dead; long live technological determinism. In Edward J. Hackett, editor, *The handbook of science and technology studies*. MIT Press, 2008.

4.4 Consequence-based Privacy Decisions: a New Way to Better Privacy Management

Zinaida Benenson (Universität Erlangen-Nürnberg, DE)
Delphine Christin (TU Darmstadt, DE)
Alexander De Luca (LMU München, DE)
Simone Fischer-Hübner (Karlstad University, SE)
Thomas Heimann (Google - München, DE)
Joachim Meyer (Tel Aviv University, IL)

4.4.1 Introduction and Motivation

An increasing number of users contribute privacy-sensitive content, such as pictures, comments, or location information, to online services. In order to protect the privacy of the users or to comply with data protection regulations, most services enable the users to customize privacy and sharing preferences. For example, this includes determining who will be authorized to access or receive and process which content and for which purposes. However, management of privacy preferences is often a fairly complex procedure that even technically savvy users often fail to understand.

Recent research shows that people would like to control their privacy and actually do so. For example, the number of Facebook users with customized privacy settings has been growing in the last years. However, users are frequently unaware of consequences resulting from their selected configuration and cannot be sure that the changes will actually have the effects they are intended to have. In addition, many users do not set or adapt privacy settings as they cannot correctly grasp the consequences of their actions. For instance, tagging a person on a photo may cause this photo to appear in searches of this person, which may be at time unwanted. Although recently some tools for granular privacy management emerged, the problem of determining all the consequences at the system level and showing them to the users in an understandable and actionable way still remains largely unsolved.

We argue that an appropriate privacy-respectful user interface should show users the consequences of making different privacy choices, rather than framing the choices only in technical terms regarding system parameters which users often do not understand and do not care about.

9

PAPER 7:

Kostenlos ist nicht kostenfrei. Oder: If you're not paying for it, you are the product

The full paper can be found at http://people.cs.kuleuven.be/~bettina.berendt/Papers/berendt_dettmar_demir_peetz_2014.pdf .

The related publications and follow-up activities are documented at http://people.cs.kuleuven.be/~bettina.berendt/Privacy-education/ .

Kostenlos ist nicht kostenfrei

oder:
"If you're not paying for it, you are the product"

Eine Unterrichtsreihe zu Datenauswertung in sozialen Netzwerken und ihren Implikationen für Privatsphäre und Demokratie

von Bettina Berendt, Gebhard Dettmar, Cihan Demir und Thomas Peetz

Eine interdisziplinäre Herausforderung

Privatsphäre und soziale Netzwerke im Unterricht

Webangebote sozialer Netzwerke wie Facebook und Twitter erfreuen sich einer großen und noch immer steigenden Beliebtheit bei Jugendlichen wie Erwachsenen. Gleichzeitig stehen sie immer stärker im Fokus einer Vielzahl von Bedenken hinsichtlich ihrer Auswirkungen auf unsere Privatsphäre (vgl. z.B. Berthold, 2010). Es geht um direkte und zeitnahe Auswirkungen, wenn etwa Facebook-Kontakte einen Wortbeitrag mit Dritten „teilen", der vom Autor eigentlich für einen kleinen Kreis gedacht war und in der nun schnell geschaffenen Öffentlichkeit als kompromittierend erfahren wird; es geht um indirekte Auswirkungen, wenn etwa früher arglos „geteilte" Partyfotos für potenzielle Arbeitgeber sichtbar sind; es geht um penetrante Werbung, die ganz offensichtlich früher besuchte Webseiten widerspiegelt. Darüber hinaus sind, wie wir spätestens seit dem Sommer 2013 wissen, all diese Inhalte auch diversen Geheimdiensten bekannt, und der Bürger wird in zunehmendem Maße „gläsern".

Die Beliebtheit gerade bei Jugendlichen – die zum Teil täglich mehrere Stunden auf Facebook verbringen – gibt besonderen Anlass zur Sorge. Eine wachsende Vielfalt von Aufklärungsmaterialien wird produziert und in verschiedener Weise im schulischen Kontext gebraucht, um die Medienkompetenz im Umgang mit diesen Plattformen zu erhöhen. In Broschüren wird versucht, zur Vorsicht zu rufen, ohne diese Absicht gleich durch den erhobenen Zeigefinger selbst zu unterlaufen. So schreibt z.B. *klicksafe*, unterstützt vom *Safer Internet Programme* der Europäischen Union (klicksafe, 2013, Punkt 4; Hervorhebungen im Original):

> Für den Schutz Deiner Privatsphäre bist Du auch selbst verantwortlich. Achte darauf, wie Du Dich im Netz zeigst!
>
> - Ein Foto darf ruhig auch mal lustig sein. Allzu *peinliche* oder *beleidigende* Fotos oder Meinungen haben in Sozialen Netzwerken aber nichts zu suchen. Sie können auch Jahre später wieder im Netz auftauchen und Dich sogar den Ausbildungsplatz kosten.
> - Überlege auch, was eine *Gruppenmitgliedschaft* über Dich aussagt. Die Gruppe „Saufen bis der Arzt kommt" ist keine gute Werbung für Dich. Hassgruppen, in denen andere gezielt beleidigt werden, gehen gar nicht.
> - Sei sorgsam mit Deinen *Profil-Daten*: Lass Anschrift, Telefon- oder ICQ-Nummern weg. Sie sind nicht nötig, wenn Du Dich mit anderen austauschst. Auch Deine private E-Mail-Adresse solltest Du nicht jedem geben.
> - Überprüfe regelmäßig Deine *Privatsphäre-Einstellungen*.
> [...]
> - Prüfe genau, wem Du freien Zugang zu Deinen *privaten* Fotos und Daten gibst. Du weißt nie, was sie mit den Informationen machen!

Dennoch scheinen solche Aufrufe weitgehend ungehört zu verhallen. Woran liegt das? Auch wenn die Rat-

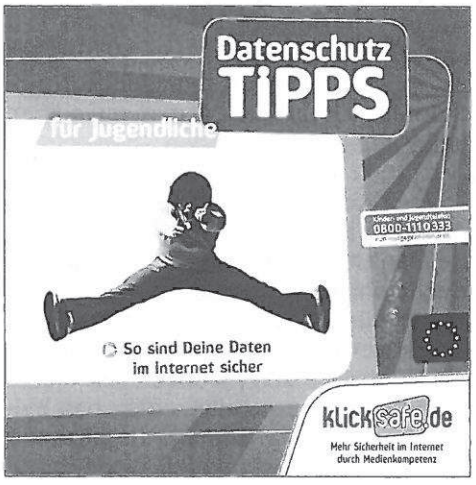


Bild 1: Die Tipps etlicher Ratgeber verhallen oft ungehört.

Quelle: klicksafe, 2013

"IF YOU'RE NOT PAYING FOR IT, YOU'RE THE PRODUCT"

TEACHING ABOUT PRIVACY IN SCHOOLS

We are developing, offering, and evaluating teaching strategies and materials about privacy in social networks and on the Web in general. We go beyond appeals to "not show your party photos to everyone on Facebook" and instead highlight implications of Big Data surveillance on fundamental rights and ultimately democracy, in order to help learners understand and reflect today's data collection and make better-informed choices. This page contains pointers to materials and activities.

The lesson series (in German) (lesson sequence, materials, 1-page overview in English)

Berendt, B., Dettmar, G., Demir, C., & Peetz, T. (2014). *Kostenlos ist nicht kostenfrei. Oder: "If you're not paying for it, you are the product"*, LOG IN 178/179: Thema "Orwell + 30" 41-56 (last version before proofs)

Berendt, B., De Paoli, S., Laing, C., Fischer-Hübner, S., Catalui, D., & Tirtea, R. (2014). *Roadmap for NIS Education.* ENISA Report.ISBN 978-92-9204-090-1

"If you're not paying for it you are the product" IniK meeting, Berlin, 4 Oct 2013. (PPT)

"If you're not paying for it, you are the product": gegevens en profielen - privacy en democratie. Talk at the SPION Educational Workshop, Brussels, 13 Feb 2014.(PPT)

Privacy-Bildung - und die Rolle der Informatik [Privacy Education – and the role of Computer Science.] Keynote at GI-FIBBB 2014, 13. GI-Tagung der Fachgruppe "Informatik-Bildung in Berlin und Brandenburg", Potsdam, Germany, 6 March 2014. (PPT)

Privatsphäre und Social Networks [Privacy and Social Networks]. Workshop at GI-FIBBB 2014, 13. GI-Tagung der Fachgruppe "Informatik-Bildung in Berlin und Brandenburg", Potsdam, Germany, 6 March 2014.

Workshop at Tagung der SH-Hill, Hamburg, Germany, 22 November 2014.

Workshop at GI-FIBBB 2014, 14. GI-Tagung der Fachgruppe "Informatik-Bildung in Berlin und Brandenburg", Berlin, Germany, 26 February 2015.

Thoughts about the need for local adaptation of teaching about privacy (*under construction*)

INTERESTED? USE OUR MATERIALS, JOIN OUR MAILING LIST, COMMENT ...

… contact us – we look forward to hearing from you!

Bettina Berendt

Dept. of Computer Science

KU Leuven, Belgium

Informatics	Economics	Society and politics
Trackers		
Profile and behavioural data		
Basic structure of data mining models (correlations in "Big Data" instead of causality)		
	Use of data by Facebook for third parties (*business models* and *customer loyalty*) → advertising	
Application of descriptive models for predicting → TIDAP (total intransparency of data analysis and processing)	*Customer segmentation* and „*weblining*" (use of data mining by third parties) → access to loans, insurance, ...	
Ex. 1: Association rule learning with Apriori		
Ex. 2: Regression analysis for prediction	Usage contexts of other third parties → access to education, work, ...?	
		The fundamental right of informational self-determination and threats to it: Chilling effects created by panoptism and TIDAP
		Plurality of opinions as a characteristic of democracy and threats to it: "Weblining" via TIDAP
	Freedom of contract vs. Other fundamental rights of participation that the state has to protect actively	

"If you're not paying for it, you are the product" :

gegevens en profielen – privacy en democratie

Bettina Berendt, KU Leuven
Gebhard Dettmar, Helmut-Schmidt-Gymnasium Hamburg

SPION Educational Workshop, 13.2.2014

Internet, comment l'utiliser ?

Hier a eu lieu dans le monde entier la journée mondiale pour un internet plus sûr. Cette journée se déroule tous les ans. Cette année, le thème était « Ensemble pour un internet meilleur ». L'objectif de cette manifestation est de faire en sorte de profiter au maximum de cet outil incroyable qu'est Internet. Mais pour y arriver, il faut se protéger de quelques méfaits !

Nous sommes tous d'accord: internet, c'est formidable! Nous trouvons tout ce que nous voulons, que ça soit pour l'école ou pour nos loisirs. Cependant, on nous dit souvent qu'il faut rester prudent. **Prudent**, oui mais de quoi? Sur internet, comme dans la vie de tous les jours, il faut se protéger. Lorsque tu traverses une rue, tu n'envisages pas de ne pas regarder à gauche et à droite si une voiture arrive? Et bien, sur internet, c'est la même chose. Il existe de petits trucs et astuces pour surfer, jouer, chatter et apprendre comme bon te semble et en toute **sécurité**.

Tout d'abord, pour surfer sans risque tout en t'amusant, tu dois informer tes **parents** de ce que tu fais quand tu surfes. Car, il y a sur internet des tas de choses qui ne te sont pas destinées et qui pourraient te filer une sacrée **frousse**! Ensuite, pour chatter en toute sécurité, la règle à suivre est d'utiliser un **pseudonyme** qui ne révèle pas si tu es un garçon ou une fille, ni ton âge, ni même ton véritable nom. Sans oublier que tu ne dois absolument **pas donner des informations** sur toi, ta famille ou tes amis. Enfin, si tu décides de te créer un profil sur un **réseau social**, n'oublie pas que c'est un journal intime ouvert à tous tes amis. Ils peuvent voir ce que tu y mets comme photos, dessins, vidéos, etc. et qui plus est, ils peuvent réagir! C'est vrai que c'est un super principe! Mais, tu dois faire attention à ne pas trop donner de détails sur toi, à ne pas poster une **photo** d'un(e) ami(e) sans son autorisation et à ne pas parler de tes sentiments trop intimes. Bref, internet est un outil fantastique qui te permet d'apprendre pleins de choses et de rester connecté avec tes amis n'importe où, n'importe quand. *Mais ne perds pas de vue ces petits conseils et n'oublie pas que tes parents sont là pour te conseiller!* (mh)

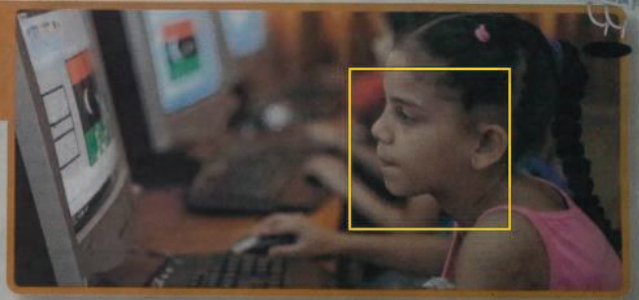
www.clicksafe.be

AFP | S. Khan

Metro, 12 februari 2014

Internet, comment l'utiliser ?

Hier a eu lieu dans le monde entier la journée mondiale pour un Internet plus sûr. Cette journée se déroule tous les ans. Cette année, le thème était « Ensemble pour un Internet meilleur ». L'objectif de cette manifestation est de faire en sorte de profiter au maximum de cet outil incroyable qu'est Internet. Mais pour y arriver, il faut se protéger de quelques méfaits !



AFP | S. Khan

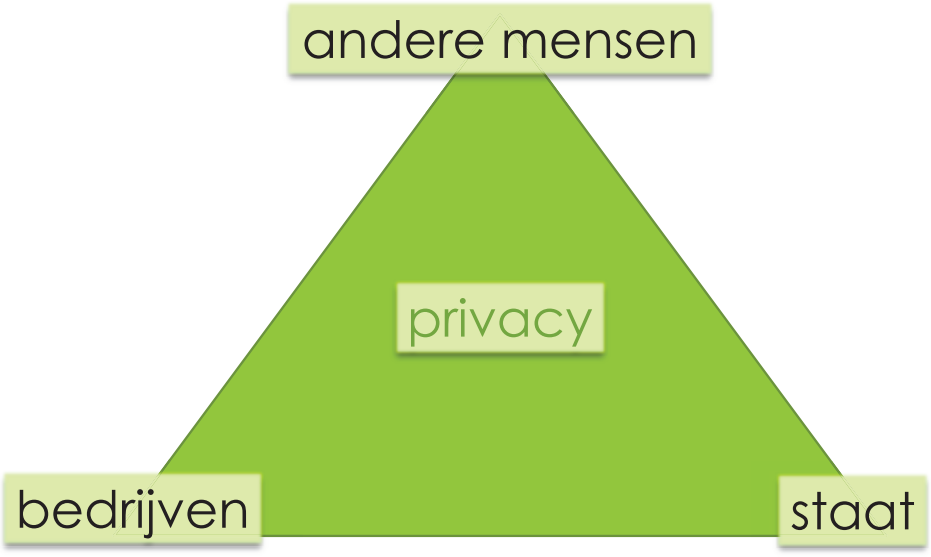
Nous sommes tous d'accord: internet, c'est formidable! Nous trouvons tout ce que nous voulons, que ça soit pour l'école ou pour nos loisirs. Cependant, on nous dit souvent qu'il faut rester prudent. Prudent, oui mais de quoi? Sur internet, comme dans la vie de tous les jours, il faut se protéger. Lorsque tu traverses une rue, tu n'envisages pas de ne pas regarder à gauche et à droite si une voiture arrive? Et bien, sur internet, c'est la même chose. Il existe de petits trucs et astuces pour surfer, jouer, chatter et apprendre comme bon te semble et en toute sécurité.

Tout d'abord, pour surfer sans risque tout en t'amusant, tu dois informer tes parents de ce que tu fais quand tu surfes. Car, il y a sur internet des tas de choses qui ne te sont pas destinées et qui pourraient te filer une sacrée frousse! Ensuite, pour chatter en toute sécurité, la règle à suivre est d'utiliser un pseudonyme qui ne révèle pas si tu es un garçon ou une fille, ni ton âge, ni même ton véritable nom. Sans oublier que tu ne dois absolument pas donner des informations sur toi, ta famille ou tes amis. Enfin, si tu décides de te créer un profil sur un réseau social, n'oublie pas que c'est un journal intime ouvert à tous tes amis. Ils peuvent voir ce que tu y mets comme photos, dessins, vidéos, etc. et qui plus est, ils peuvent réagir! C'est vrai que c'est un super principe! Mais, tu dois faire attention à ne pas trop donner de détails sur toi, à ne pas poster une photo d'un(e) ami(e) sans son autorisation et à ne pas parler de tes sentiments trop intimes. Bref, internet est un outil fantastique qui te permet d'apprendre pleins de choses et de rester connecté avec tes amis n'importe où, n'importe quand. Mais ne perds pas de vue ces petits conseils et n'oublie pas que tes parents sont là pour te conseiller!

(mh)

www.clicksafe.be

Privacy vis-à-vis wie?



Gegevensverzamelingen en informationele zelfbeschikking

- "[The data] can also [...] be combined with other data collections to yield a partial or basically complete picture of someone's personality,
- without the concerned person being able to sufficiently control the correctness and use of this ensemble. [...]
- The right to informational self-determination would be incompatible with a social order and an underlying legal order in which citizens can no longer know who knows what and when under which circumstances about them.

Het grondrecht op informationele zelfbeschikking en democratie (1)

- Iedereen die er niet zeker van kan zijn dat gegevens over maatschappelijk afwijkend gedrag voor langere tijd worden geregistreerd en kunnen worden gebruikt op een manier waarvan hij niets weet, zal proberen om dat gedrag niet te vertonen.
- Dat is in strijd met de elementaire functie van zelfbeschikking in een democratische samenleving waarin de burgers de mogelijkheid moeten hebben om deel te nemen aan het maatschappelijke en politieke leven zonder risico te lopen op een voor hen ondoorzichtige manier te worden geregistreerd."

Bundesverfassungsgericht[grondwettelijk hof van Duitsland], 1983

gegevensbescherming

Het grondrecht op informationele zelfbestemming en democratie (2)

- 28 januari 2014: **Liga vraagt Parlementaire Onderzoekscommissie over NSA**
- **Naar Duits voorbeeld vraagt de Liga voor Mensenrechten de oprichting van een parlementaire onderzoekscommissie om de betrokkenheid van de binnenlandse inlichtingendiensten bij de grootschalige spionagepraktijken van de Amerikaanse NSA tegen het licht te houden. Hiertoe werd een verzoek gericht tot Kamervoorzitter André Flahaut.**
- [...] De Liga wijst op het belang van een democratische controle op de samenwerking van binnenlandse en buitenlandse inlichtingendiensten en van mogelijke maatregelen om onaanvaardbare en buitensporige spionageactiviteiten te voorkomen en aan banden te leggen. "De bescherming van de persoonlijke levenssfeer is een kostbaar goed en moet met alle mogelijke democratische middelen worden nagestreefd en gehandhaafd, zegt Paul Pataer van de Liga voor Mensenrechten."
- Betrouwbare bronnen bevestigen dat verschillende bedrijven, vooral actief in de IT-sector, meewerken met de NSA in het verzamelen van contactgegevens.
- Het is nauwelijks denkbaar dat de eigen binnenlandse veiligheids- en inlichtingendiensten niet op de hoogte zijn van die activiteiten en het is niet ondenkbaar dat diezelfde diensten bijstand hebben verleend en verlenen aan de NSA bij het ontplooien van haar activiteiten.

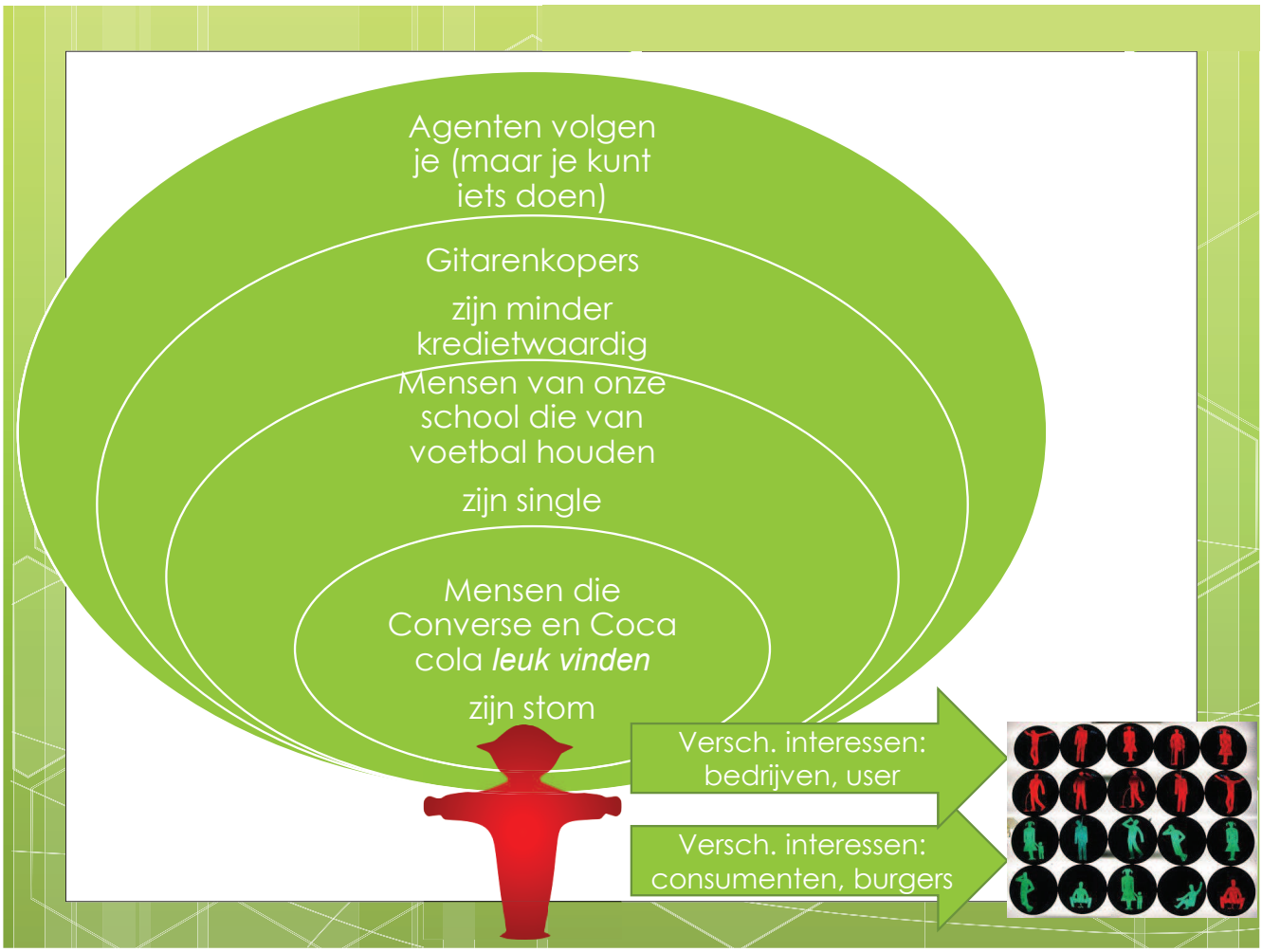
De lessenreeks: overzicht

- Doel: kennis, besef van belang van IZ, controle
- Begin 3e graad ASO
- 10 * 2u
- Interdisciplinair: tussen „politiek, maatschappij, economie" en „informatica"
- Op de volgende 3 pagina's:
  - thema's
  - didactische vormen
  - "betrokkenheid / verontwaardiging creëren"

Informatica	Economie	Maatschappij en politiek
Trackers		
Profiel- en gedragsgegevens		
Basisstructuur van data mining modellen (correlaties in „Big Data“ i.p.v. causaliteit)		
	Gebruik van gegevens door Facebook voor derden (*business models* en *customer loyalty*) → advertenties	
Toepassing van descriptieve modellen voor voorspelling → VIGAV (volledige intransparantie van gegevensanalyse en –verwerking)	*Customer segmentation* en „*weblining*“ (gebruik van data mining door derden) → toegang tot credieten, verzekeringen, ...	
Vb. 1: associatieregels leren met Apriori		
Vb. 2: regressieanalyse voor voorspelling	Gebruikscontexten van andere derden → toegang tot opleiding, werk, ...?	
		Het grondrecht op informationele zelfbestemming en zijn bedreigingen: Chilling-effecten door panoptisme en VIGAV
		Pluraliteit van meningen als kenmerk van democratie en zijn bedreigingen: „Weblining“ door VIGAV
	Contractvrijheid vs. andere (deelname)grondrechten die de staat positief moet beschermen	

Informatica	Economie	Maatschappij en politiek
Trackers		
Profiel- en gedragsgegevens		
Basisstructuur van data mining modellen (correlaties in „Big Data“ i.p.v. causaliteit)		
	Gebruik van gegevens door Facebook voor derden (*business models* en *customer loyalty*) → advertenties	
Toepassing van descriptieve modellen voor voorspelling → VIGAV (volledige intransparantie van gegevensanalyse en –verwerking)	*Customer segmentation* en „*weblining*“ (gebruik van data mining door derden) → toegang tot credieten, verzekeringen, ...	
Vb. 1: associatieregels leren met Apriori		
Vb. 2: regressieanalyse voor voorspelling	Gebruikscontexten van andere derden → toegang tot opleiding, werk, ...?	
		Het grondrecht op informationele zelfbestemming en zijn bedreigingen: Chilling-effecten door panoptisme en VIGAV
		Pluraliteit van meningen als kenmerk van democratie en zijn bedreigingen: „Weblining“ door VIGAV
	Contractvrijheid vs. andere (deelname)grondrechten die de staat positief moet beschermen	

Informatica	Economie	Maatschappij en politiek
Tekst (geschreven voor het SPION *Privacy Manual*) + software tools ter bescherming tegen gegevensverzameling		
Tekst (Website voor een breed publiek)		
Tekst (kwaliteitskrant)		
	Tekst (voor en seminaar; Facebook's Data Use Policy)	
Tekst (kwaliteitskrant)	(Tekst zie links) Rollenspel	
Web API (Facebook) + data mining algoritme		
Data mining online tool (Preference Tool: "Predicting personality from Faceb. *Likes*")	Documentatie rond het tool, wetenschappelijk artikel - psychologie	
		Teksten(rechtbank ordeel; wetenschappelijk artikel - rechten)
		Tekst (wetenschappelijk artikel - rechten)
		Rollenspel



Besluit

- Succes:
  - 1 ½ keer doorgevoerd
  - veranderingen in houding en gedrag
- Hoofduitdagingen:
  - In welk vak?
  - Meer ervaringen verzamelen
    o.a. in verschillende landen/culturen

→ Doe mee in Vlaanderen ☺